Input-Output Stability of First-Order Optimization Algorithms: A Passivity Approach

Sepehr Moalemi James Richard Forbes

McGill

DECAR

{ ISMP 2024 } 25th International Symposium on Mathematical Programming

July 21-26 MONTRÉAL

July 24, 2024

A Classic Optimization Problem

Consider the unconstrained optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function satisfying:

1. L-Lipschitz gradient: $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, where $L \in \mathbb{R}_{>0}$.
2. m-strongly convex: $g(\mathbf{x}) \coloneqq f(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$ is convex and $m \in \mathbb{R}_{>0}$.

Remark

*Since f is m-strongly convex, it has a **unique global minimizer** $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$.*

First-Order Optimization Algorithms

- Gradient Descent (GD):

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

- Polyak's Heavy-Ball (HB):

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

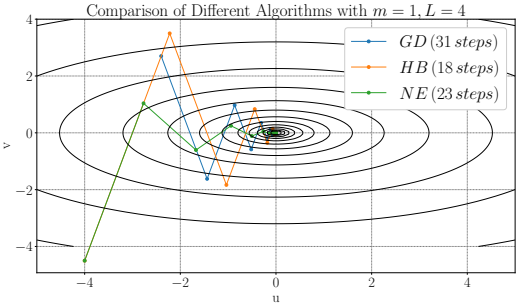
- Nesterov's Accelerated Gradient (NE):

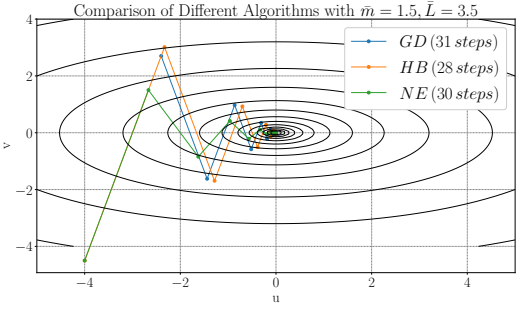
$$\mathbf{x}_{k+1} = (1+\beta)\mathbf{x}_k - \beta \mathbf{x}_{k-1} - \alpha \nabla f((1+\beta)\mathbf{x}_k - \beta \mathbf{x}_{k-1}).$$

Optimal Tuning Example

Table 1: Optimal Tuning Parameters for the Class of Convex Quadratic Functions ($\kappa = \frac{L}{m}$).

	Optimal Parameter	
Method	α	β
GD	$\frac{2}{L+m}$	
HB	$\frac{4}{(\sqrt{L}+\sqrt{m})^2}$	$\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$
NE	$\frac{4}{3L+m}$	$\frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$





Goal

- Relax the assumptions on the function f by considering the class of functions that have **sector bounded gradients**.
- Establish regions of stability in the presence of **model uncertainty** (e.g., the true value of m is unknown).
- Characterize **passivity** properties of first-order optimization algorithms.

Outline

- Control Theory Meets Optimization
- Introduction to Passivity-Based Control
- Passivity of Gradient Descent
- Ongoing and Future Work

Outline

- Control Theory Meets Optimization
- Introduction to Passivity-Based Control
- Passivity of Gradient Descent
- Ongoing and Future Work

Control Theory Meets Optimization

- *Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints* [Lessard, Recht, Packard, 2015].
- *Control Interpretations for First-Order Optimization Methods* [Hu, Lessard, 2017].
- *The Fastest Known Globally Convergent First-Order Method for Minimizing Strongly Convex Functions* [Van Scoy, Freeman, Lynch, 2018].
- *The Analysis of Optimization Algorithms, A Dissipativity Approach* [Lessard, 2022].
- *A Tutorial on a Lyapunov-Based Approach to the Analysis of Iterative Optimization Algorithms* [Van Scoy, Lessard, 2023].

Linear Dynamical System

A linear dynamical system can be written as a set of recursive equations of the form

$$\mathbf{y}_k \longleftarrow \boxed{\mathcal{G}: \left\{ \begin{array}{r} \boldsymbol{\xi}_{k+1} = \mathbf{A}\boldsymbol{\xi}_k + \mathbf{B}\mathbf{u}_k \\ \mathbf{y}_k = \mathbf{C}\boldsymbol{\xi}_k + \mathbf{D}\mathbf{u}_k \end{array} \right\}} \longleftarrow \mathbf{u}_k$$

where

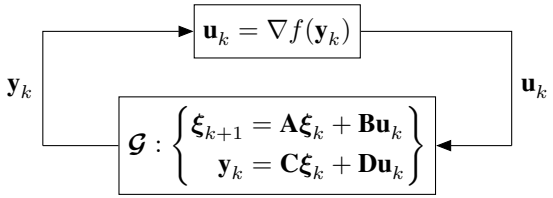
- $\boldsymbol{\xi}_k \in \mathbb{R}^d$ is the **state**,
- $\mathbf{u}_k \in \mathbb{R}^m$ is the **input**,
- $\mathbf{y}_k \in \mathbb{R}^n$ is the **output**.

More compactly, the dynamical system can be represented using the notation

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right].$$

Special Case of the Lur'e Problem

Consider this **linear time-invariant** dynamical system, $\mathcal{G}$, connected in feedback with a **static memory-less nonlinearity**, ∇f, given by



For the special case of $\mathbf{A} = \mathbf{1}$, $\mathbf{B} = -\alpha\mathbf{1}$, $\mathbf{C} = \mathbf{1}$, and $\mathbf{D} = \mathbf{0}$, it follows that

$$\mathcal{G}: \begin{cases} \boldsymbol{\xi}_{k+1} = \boldsymbol{\xi}_k - \alpha\mathbf{u}_k \\ \mathbf{y}_k = \boldsymbol{\xi}_k \\ \mathbf{u}_k = \nabla f(\mathbf{y}_k) \end{cases} \quad \Longleftrightarrow \quad \boldsymbol{\xi}_{k+1} = \boldsymbol{\xi}_k - \alpha\nabla f(\boldsymbol{\xi}_k) \text{ (GD)}$$

Control Interpretation of First-Order Optimization Algorithms [Lessard, Recht, Packard, 2015]

- Gradient Descent (GD):

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k) \iff \left[\begin{array}{c|c} \mathbf{1} & -\alpha \mathbf{1} \\ \hline \mathbf{1} & \mathbf{0} \end{array}\right].$$

- Polyak's Heavy-Ball (HB):

$$\mathbf{x}_{k+1} = (1+\beta)\mathbf{x}_k - \beta \mathbf{x}_{k-1} - \alpha \nabla f(\mathbf{x}_k) \iff \left[\begin{array}{cc|c} (1+\beta)\mathbf{1} & -\beta \mathbf{1} & -\alpha \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{1} & \mathbf{0} & \mathbf{0} \end{array}\right].$$

- Nesterov's Accelerated Gradient (NE):

$$\mathbf{x}_{k+1} = (1+\beta)\mathbf{x}_k - \beta \mathbf{x}_{k-1} - \alpha \nabla f((1+\beta)\mathbf{x}_k - \beta \mathbf{x}_{k-1}) \iff \left[\begin{array}{cc|c} (1+\beta)\mathbf{1} & -\beta \mathbf{1} & -\alpha \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \hline (1+\beta)\mathbf{1} & -\beta \mathbf{1} & \mathbf{0} \end{array}\right].$$

Outline

- Control Theory Meets Optimization
- Introduction to Passivity-Based Control
- Passivity of Gradient Descent
- Ongoing and Future Work

Introduction to Passivity

Definition (Passivity in Discrete Time [Desoer, Vidyasagar, 1975])

Consider a square system with input $\mathbf{u} \in \ell_{2e}$ and output $\mathbf{y} \in \ell_{2e}$ mapped through the operator $\mathcal{G} : \ell_{2e} \to \ell_{2e}$. The system $\mathcal{G}$ is

- **passive** if $\exists \beta \in \mathbb{R}_{\leq 0}$ s.t.

$$\langle \mathbf{y}, \mathbf{u} \rangle_T \geq \beta, \qquad \forall \mathbf{u} \in \ell_{2e}, \forall T \in \mathbb{Z}^+, \tag{2}$$

- **very strictly passive (VSP)** if $\exists \delta \in \mathbb{R}_{>0}$, $\exists \varepsilon \in \mathbb{R}_{>0}$, and $\exists \beta \in \mathbb{R}_{\leq 0}$ s.t.

$$\langle \mathbf{y}, \mathbf{u} \rangle_T \geq \beta + \delta \|\mathbf{u}\|_{2T}^2 + \varepsilon \|\mathbf{y}\|_{2T}^2, \qquad \forall \mathbf{u} \in \ell_{2e}, \forall T \in \mathbb{Z}^+. \tag{3}$$

Passivity Theorem

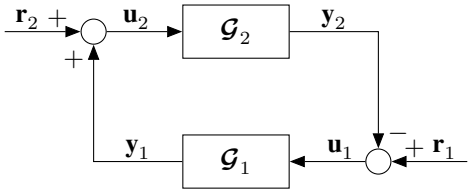


Figure 1: The negative feedback interconnection of two systems $\mathcal{G}_1$ and $\mathcal{G}_2$.

Theorem (Passivity Theorem in Discrete Time [Desoer, Vidyasagar, 1975])

*Consider the negative feedback interconnection of $\mathcal{G}_1 : \ell_{2e} \to \ell_{2e}$ and $\mathcal{G}_2 : \ell_{2e} \to \ell_{2e}$ in Figure 1. Provided $\mathcal{G}_1$ is **passive** and $\mathcal{G}_2$ is **VSP**, the negative feedback interconnection is ℓ_2-stable.*

Characterization of Passivity

Lemma (Hitz and Anderson, 1969)

A dynamical system of the form

$$\begin{cases} \boldsymbol{\xi}_{k+1} = \mathbf{A}\boldsymbol{\xi}_k + \mathbf{B}\mathbf{u}_k, \\ \mathbf{y}_k = \mathbf{C}\boldsymbol{\xi}_k + \mathbf{D}\mathbf{u}_k, \end{cases}$$

*is **passive** with respect to the storage function* $V_k = \frac{1}{2}\mathbf{x}_k^\mathsf{T}\mathbf{P}\mathbf{x}_k$*, if and only if, there exists a* $\mathbf{P} = \mathbf{P}^\mathsf{T} \succ 0$ *such that*

$$\mathbf{A}^\mathsf{T}\mathbf{P}\mathbf{A} - \mathbf{P} \preceq 0, \tag{4a}$$

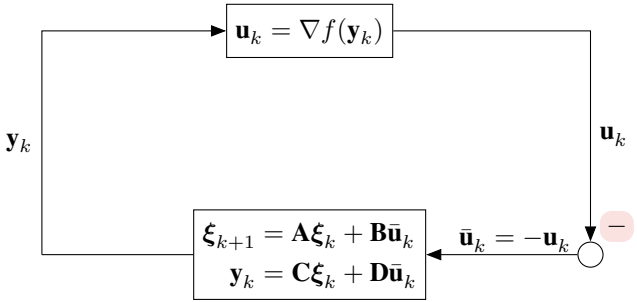
$$\mathbf{B}^\mathsf{T}\mathbf{P}\mathbf{A} = \mathbf{C}, \tag{4b}$$

$$\mathbf{B}^\mathsf{T}\mathbf{P}\mathbf{B} - \left(\mathbf{D}^\mathsf{T} + \mathbf{D}\right) \preceq 0. \tag{4c}$$

Outline

- Control Theory Meets Optimization
- Introduction to Passivity-Based Control
- Passivity of Gradient Descent
- Ongoing and Future Work

Revisit the Lur'e Problem

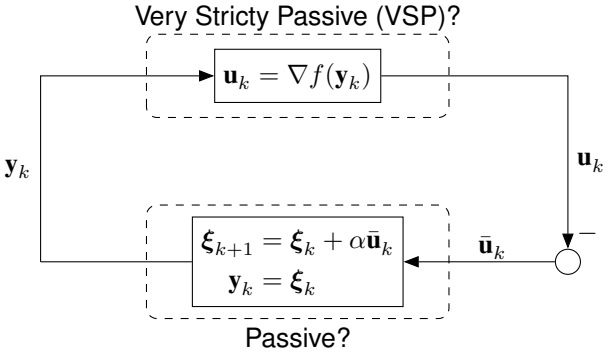


GD: $\left[\begin{array}{c|c} \mathbf{1} & \alpha\mathbf{1} \\ \hline \mathbf{1} & \mathbf{0} \end{array}\right]$

HB: $\left[\begin{array}{cc|c} (1+\beta)\mathbf{1} & -\beta\mathbf{1} & \alpha\mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{1} & \mathbf{0} & \mathbf{0} \end{array}\right]$

NE: $\left[\begin{array}{cc|c} (1+\beta)\mathbf{1} & -\beta\mathbf{1} & \alpha\mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \hline (1+\beta)\mathbf{1} & -\beta\mathbf{1} & \mathbf{0} \end{array}\right]$

Goal: Passivity of Gradient Descent



Review (Passivity Theorem in Discrete Time)

Consider the negative feedback interconnection of $\mathcal{G}_1 : \ell_{2e} \to \ell_{2e}$ and $\mathcal{G}_2 : \ell_{2e} \to \ell_{2e}$ in Figure 1. Provided $\mathcal{G}_1$ is passive and $\mathcal{G}_2$ is VSP, the negative feedback interconnection is ℓ_2-stable.

Apply Lemma 3 to Gradient Descent

GD: $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right] = \left[\begin{array}{c|c} 1 & \alpha \\ \hline 1 & 0 \end{array}\right]$

Review (Hitz and Anderson, 1969)

$$\mathbf{A}^{\mathsf{T}}\mathbf{P}\mathbf{A} - \mathbf{P} \preceq 0, \tag{4a}$$

$$\mathbf{B}^{\mathsf{T}}\mathbf{P}\mathbf{A} = \mathbf{C}, \tag{4b}$$

$$\mathbf{B}^{\mathsf{T}}\mathbf{P}\mathbf{B} - \left(\mathbf{D}^{\mathsf{T}} + \mathbf{D}\right) \preceq 0. \tag{4c}$$

(4a) $\implies P - P = 0 \leq 0$ ✓

(4b) $\implies \alpha P = 1 \implies P = \frac{1}{\alpha} > 0,$ for $\alpha \in \mathbb{R}_{>0}$ ✓

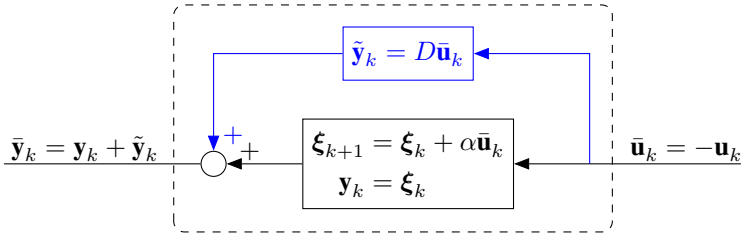
(4c) $\implies$ $\alpha^2 P = \alpha \not\leq 0$

Remark (Byrnes and Lin, 1994)

Passivity *and losslessness of* ***discrete-time*** *systems having outputs independent of* $\mathbf{u}_k$ *(i.e.,* $\mathbf{D} = \mathbf{0}$*)* ***cannot*** *be studied.*

What if we add D?

$$\bar{\mathcal{G}}: \left\{ \begin{aligned} \boldsymbol{\xi}_{k+1} &= \boldsymbol{\xi}_k + \alpha \bar{\mathbf{u}}_k \\ \bar{\mathbf{y}}_k &= \boldsymbol{\xi}_k + D\bar{\mathbf{u}}_k \end{aligned} \right\}$$



Question: What is the minimum D required for Passivity?

Modified GD: $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right] = \left[\begin{array}{c|c} 1 & \alpha \\ \hline 1 & D \end{array}\right]$

Review (Hitz and Anderson, 1969)

$$\mathbf{A}^{\mathsf{T}}\mathbf{P}\mathbf{A} - \mathbf{P} \preceq 0, \tag{4a}$$

$$\mathbf{B}^{\mathsf{T}}\mathbf{P}\mathbf{A} = \mathbf{C}, \tag{4b}$$

$$\mathbf{B}^{\mathsf{T}}\mathbf{P}\mathbf{B} - \left(\mathbf{D}^{\mathsf{T}} + \mathbf{D}\right) \preceq 0. \tag{4c}$$

(4a) $\implies P - P = 0 \leq 0$ ✓

(4b) $\implies \alpha P = 1 \implies P = \frac{1}{\alpha} > 0,$ for $\alpha \in \mathbb{R}_{>0}$ ✓

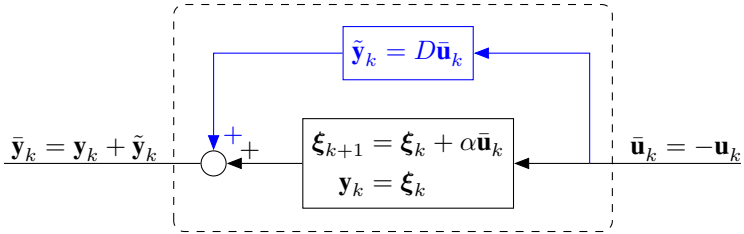
(4c) $\implies \alpha^2 P - 2D \leq 0 \implies \frac{\alpha}{2} \leq D$ ✓

Remark

The necessary and sufficient condition for $\left[\begin{array}{c|c} 1 & \alpha \\ \hline 1 & D \end{array}\right]$ *to be passive is* $0 < \frac{\alpha}{2} \leq D$.

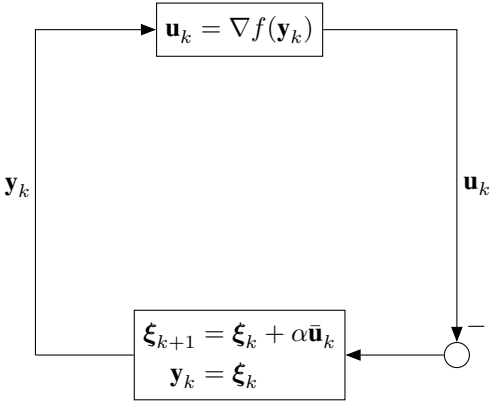
Recap: What if we add D?

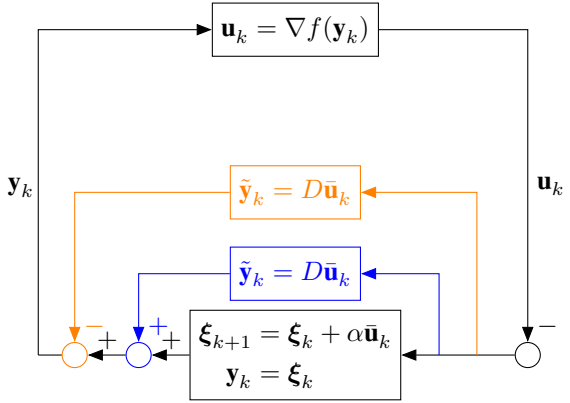
$$\bar{\mathcal{G}}: \left\{ \begin{array}{l} \boldsymbol{\xi}_{k+1} = \boldsymbol{\xi}_k + \alpha \bar{\mathbf{u}}_k \\ \bar{\mathbf{y}}_k = \boldsymbol{\xi}_k + D\bar{\mathbf{u}}_k \end{array} \right\}$$



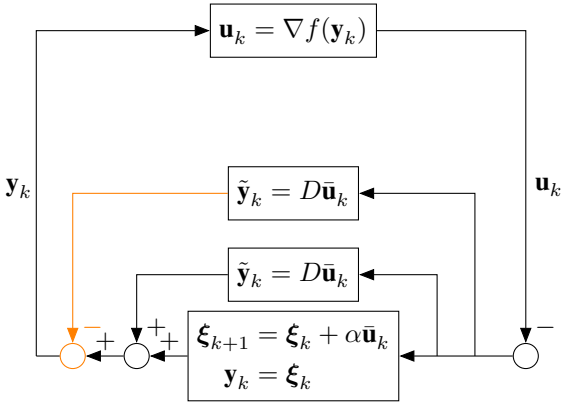
Shown to be Passive iff $\frac{\alpha}{2} \leq D$

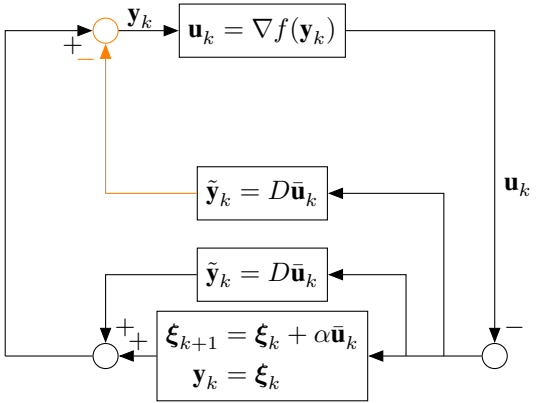
Loop Transformation



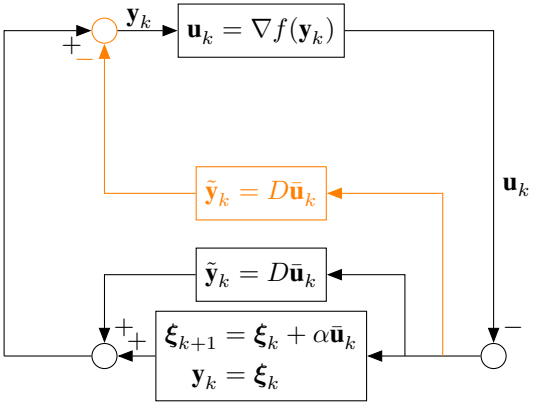


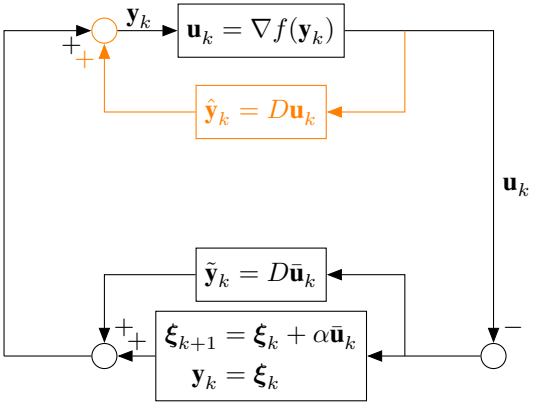
Loop Transformation (Cont'd)



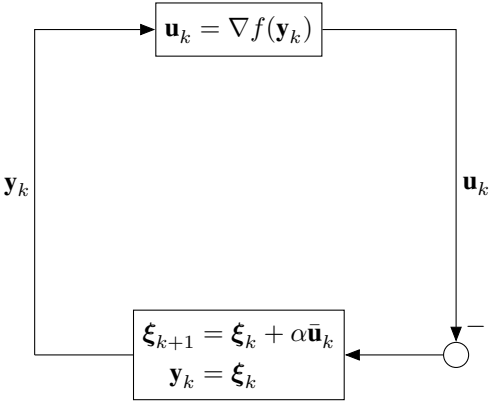


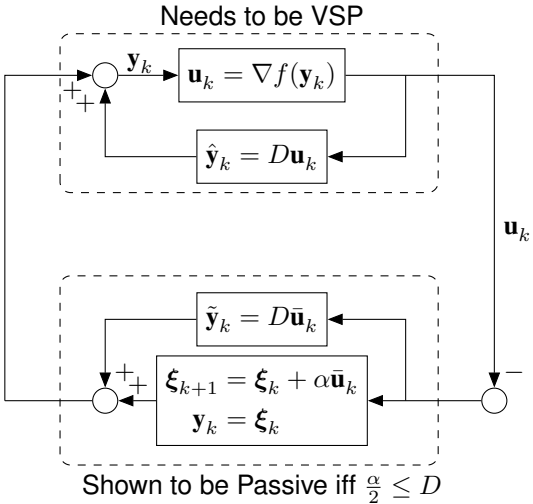
Loop Transformation (Cont'd)





Loop Transformation (Cont'd)





Revisit Assumptions on f

Definition (Function class $\mathcal{F}_{m,L}$)

The set of continuously differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ that are m-strongly convex and L-smooth is denoted by $\mathcal{F}_{m,L}$. Additionally, $\kappa = \frac{L}{m}$ is called the condition number of f.

Remark (Co-coercivity)

Suppose $f \in \mathcal{F}_{m,L}$, then the function $g(\mathbf{x}) = f(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$ is convex and $(L-m)$-smooth. The co-coercivity of ∇g can be written as

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^{\mathsf{T}}(\mathbf{x} - \mathbf{y}) \geq \frac{mL}{m+L}\|\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{m+L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2, \tag{5}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Control Interpretation of Co-coercivity

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\mathsf{T}(\mathbf{x}-\mathbf{y}) \geq \frac{mL}{m+L}\|\mathbf{x}-\mathbf{y}\|_2^2 + \frac{1}{m+L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Consider the pair of points $(\mathbf{x}_k, \mathbf{x}^*)$ for $k \in [0, 1, \ldots, T-1]$, where $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$. The co-coercivity inequality can be written as

$$\nabla f(\mathbf{x}_k)^\mathsf{T}\left(\mathbf{x}_k - \mathbf{x}^*\right) \geq \frac{mL}{m+L}\left\|\mathbf{x}_k - \mathbf{x}^*\right\|_2^2 + \frac{1}{m+L}\|\nabla f(\mathbf{x}_k)\|_2^2. \tag{6}$$

Summing (6) over all points yields

$$\sum_{k=0}^{T-1} \nabla f(\mathbf{x}_k)^\mathsf{T}\left(\mathbf{x}_k - \mathbf{x}^*\right) = \left\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \right\rangle_T \geq \underbrace{\frac{mL}{m+L}}_{\delta}\left\|\mathbf{x} - \mathbf{x}^*\right\|_{2T}^2 + \underbrace{\frac{1}{m+L}}_{\varepsilon}\|\nabla f(\mathbf{x})\|_{2T}^2. \tag{7}$$

Remark

*Suppose $f \in \mathcal{F}_{m,L}$, then ∇f is **VSP** with $\delta = \frac{mL}{m+L}$ and $\varepsilon = \frac{1}{m+L}$.*

Sector Bounded Condition

Consider the scalar function $u_k = \phi(y_k)$ where the function $\phi : \mathbb{R} \to \mathbb{R}$ satisfies $\phi(0) = 0$. The function ϕ is said to be sector bounded if it lies in a sector formed by two lines with slopes m and L such that $m \leq L$. Therefore,

$$my_k \leq u_k \leq Ly_k \qquad \Longleftrightarrow \qquad 0 \leq (Ly_k - u_k)^{\mathsf{T}}(u_k - my_k). \tag{8}$$

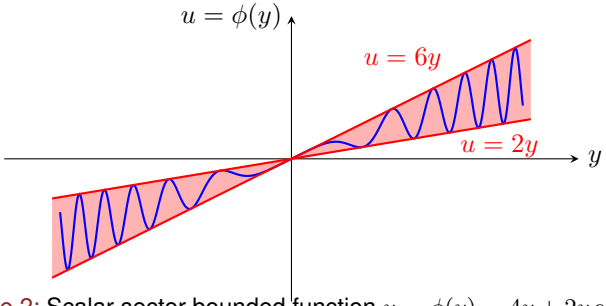


Figure 2: Scalar sector bounded function $u = \phi(y) = 4y + 2y\cos(y^2)$.

Generalized Sector Bound

Definition (Function class $\mathcal{S}_{m,L}$)

Given the constants $0 < m \leq L$, the set of twice differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ that satisfy

$$\big(m\big(\mathbf{x}-\mathbf{x}^*\big)-\nabla f(\mathbf{x})\big)^{\mathsf{T}}\big(L\big(\mathbf{x}-\mathbf{x}^*\big)-\nabla f(\mathbf{x})\big)\leq 0, \qquad \forall \mathbf{x}\in\mathbb{R}^n, \tag{9}$$

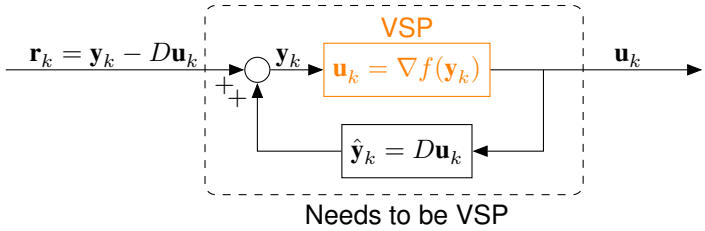
where $\mathbf{x}^*$ is the unique minimizer of f, is denoted by $\mathcal{S}_{m,L}$.

From co-coercivity: $$\nabla f(\mathbf{x}_k)^{\mathsf{T}}(\mathbf{x}_k-\mathbf{x}^*)\geq \frac{mL}{m+L}\|\mathbf{x}_k-\mathbf{x}^*\|_2^2+\frac{1}{m+L}\|\nabla f(\mathbf{x}_k)\|_2^2. \tag{10}$$

Remark

Expanding (9) leads to the (10) and therefore, $\mathcal{F}_{m,L}\subseteq\mathcal{S}_{m,L}$.

Is the Feedback Interconnection VSP?



We need to show: $\exists \bar{\delta}, \bar{\varepsilon} \in \mathbb{R}_{>0}$ such that $\langle \mathbf{y} - D\mathbf{u}, \mathbf{u} \rangle_T \geq \bar{\delta} \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \bar{\varepsilon} \|\mathbf{u}\|_{2T}^2$.

Is the Feedback Interconnection VSP?

We know: $\langle \mathbf{y}, \mathbf{u} \rangle_T \geq \delta \|\mathbf{y}\|_{2T}^2 + \varepsilon \|\mathbf{u}\|_{2T}^2$.

We need to show: $\exists \bar{\delta}, \bar{\varepsilon} \in \mathbb{R}_{>0}$ such that $\langle \mathbf{y} - D\mathbf{u}, \mathbf{u} \rangle_T \geq \bar{\delta} \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \bar{\varepsilon} \|\mathbf{u}\|_{2T}^2$.

Proof:

$$
\begin{aligned}
\langle \mathbf{y}, \mathbf{u} \rangle_T &\geq \delta \|\mathbf{y}\|_{2T}^2 + \varepsilon \|\mathbf{u}\|_{2T}^2 \\
&= \delta \|\mathbf{y}\|_{2T}^2 + \varepsilon \|\mathbf{u}\|_{2T}^2 + 2\delta D \langle \mathbf{y}, \mathbf{u} \rangle_T - 2\delta D \langle \mathbf{y}, \mathbf{u} \rangle_T + \delta D^2 \|\mathbf{u}\|_{2T}^2 - \delta D^2 \|\mathbf{u}\|_{2T}^2 \\
&\geq \delta \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \left(\varepsilon + 2\delta\varepsilon D - \delta D^2 \right) \|\mathbf{u}\|_{2T}^2.
\end{aligned}
$$

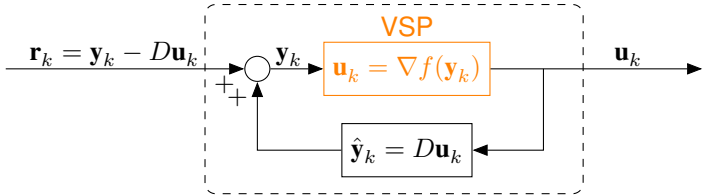
Therefore,

$$
\langle \mathbf{y} - D\mathbf{u}, \mathbf{u} \rangle_T \geq \underbrace{\delta}_{\bar{\delta}} \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \underbrace{\left(\varepsilon + (2\delta\varepsilon - 1)D - \delta D^2 \right)}_{\bar{\varepsilon}} \|\mathbf{u}\|_{2T}^2.
$$

Remark

As $D \to 0$ the VSP property of ∇f is recovered.

Proof (Contd.)

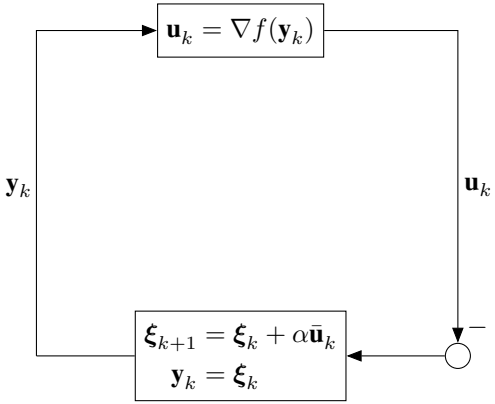


We need to show: $\exists \bar{\delta}, \bar{\varepsilon} \in \mathbb{R}_{>0}$ such that $\langle \mathbf{y} - D\mathbf{u}, \mathbf{u} \rangle_T \geq \bar{\delta} \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \bar{\varepsilon} \|\mathbf{u}\|_{2T}^2$.

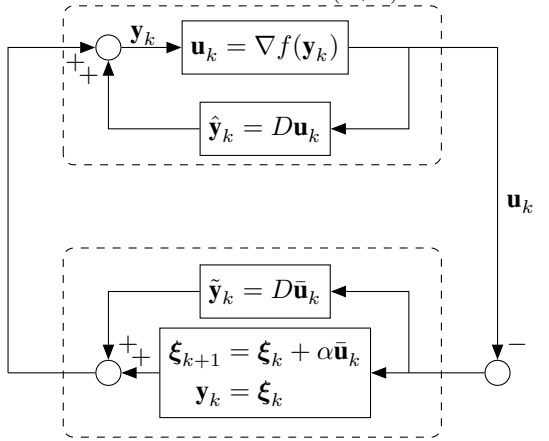
$$\langle \mathbf{y} - D\mathbf{u}, \mathbf{u} \rangle_T \geq \underbrace{\delta}_{\bar{\delta}} \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \underbrace{\left(\varepsilon + (2\delta\varepsilon - 1)D - \delta D^2\right)}_{\bar{\varepsilon}} \|\mathbf{u}\|_{2T}^2. \tag{11}$$

- $\bar{\delta} = \delta = \frac{mL}{m+L} \in \mathbb{R}_{>0}$ ✓
- $\bar{\varepsilon} = \varepsilon + (2\delta\varepsilon - 1)D - \delta D^2 \in \mathbb{R}_{>0}$ for $D \in \left[0, \frac{\sqrt{L^4 + 4L^3 m + 10L^2 m^2 + 4Lm^3 + m^4} - L^2 - m^2}{2Lm(L+m)}\right)$ ✓

Recap



VSP for $D < \frac{\sqrt{L^4+4L^3m+10L^2m^2+4Lm^3+m^4}-L^2-m^2}{2Lm(L+m)}$



Passive iff $\frac{\alpha}{2} \leq D$

ℓ_2-Stability of Gradient Descent using the Passivity Theorem

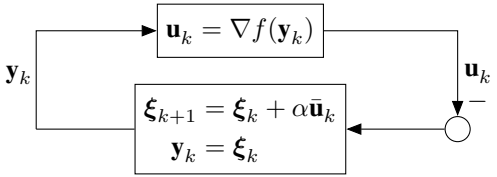


Figure 3: Gradient descent as a negative feedback loop.

Theorem (ℓ_2-Stability of Gradient Descent)

Consider the gradient descent algorithm shown in Figure 3, where $\alpha \in \mathbb{R}_{>0}$ is the step-size and $f \in \mathcal{S}_{m,L}$ with $0 < m \leq L$. Provided the step-size satisfies

$$\alpha < \frac{\sqrt{L^4 + 4L^3m + 10L^2m^2 + 4Lm^3 + m^4} - L^2 - m^2}{Lm(L+m)},$$

the negative feedback interconnection is ℓ_2-stable.

Summary

- It is possible to relax the assumptions on the function f by considering the class of functions that have sector bounded gradients.
- GD can be analyzed within the framework of passivity theory by including a feedthrough term using a loop transformation.
- At the face of model uncertainty, the passivity theorem can be used to ensure the ℓ_2-stability of GD as long as the VSP and passive properties of the negative feedback interconnection are preserved.
- To analyze the stability of other first-order optimization algorithms, we only need to find an appropriate feedthrough term to render the algorithm passive.

Outline

- Control Theory Meets Optimization
- Introduction to Passivity-Based Control
- Passivity of Gradient Descent
- Ongoing and Future Work

Ongoing and Future Work

- QSR-dissipativity is a generalization of passivity:

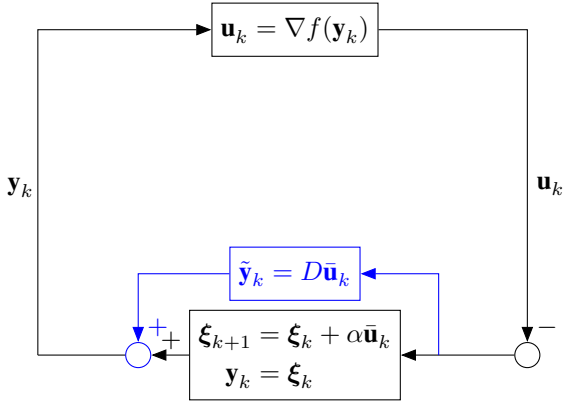
$$\langle \mathbf{y}, \mathbf{Q}\mathbf{y} \rangle_T + 2 \langle \mathbf{y}, \mathbf{S}\mathbf{u} \rangle_T + \langle \mathbf{u}, \mathbf{R}\mathbf{u} \rangle_T \geq V(\mathbf{x}(T)) - V(\mathbf{x}(0)).$$

  A few special cases of QSR-dissipativity are
  - *passive* if $\mathbf{Q} = \mathbf{0}$, $\mathbf{S} = \frac{1}{2}\mathbf{1}$, and $\mathbf{R} = \mathbf{0}$,
  - *very strictly passive (VSP)* if $\mathbf{Q} = -\varepsilon\mathbf{1}$, $\mathbf{S} = \frac{1}{2}\mathbf{1}$, and $\mathbf{R} = -\delta\mathbf{1}$, where $\delta, \varepsilon \in \mathbb{R}_{>0}$, and
  - *conic* if $\mathbf{Q} = -\mathbf{1}$, $\mathbf{S} = \frac{a+b}{2}\mathbf{1} = c\mathbf{1}$, and $\mathbf{R} = -ab\mathbf{1} = (r^2 - c^2)\mathbf{1}$, where $a, b \in \mathbb{R}$ represent lower and upper conic bounds, while $c \in \mathbb{R}$ and $r \in \mathbb{R}_{>0}$ represent the center and radius of the conic sector, respectively.

Remark

GD is QSR-dissipative with $\mathbf{Q} = 0$, $\mathbf{S} = \frac{1}{2}$, and $\mathbf{R} = \frac{\alpha}{2}$ so that is why adding a feedthrough of $D = \frac{\alpha}{2}$ pushes GD to be passive (i.e., $\mathbf{R} = \mathbf{0}$).

Ongoing and Future Work



- What if we dont do a loop transformation? what are the properties of this new modified GD?
  - From passivity theorem we already know it is ℓ_2-stable.

Questions?

sepehr.moalemi@mail.mcgill.ca james.richard.forbes@mcgill.ca

Natural Sciences and Engineering Research Council of Canada

Conseil de recherches en sciences naturelles et en génie du Canada

Canada

Backup Slides

Is the Feedback Interconnection VSP?

We know: $\langle \mathbf{y}, \mathbf{u} \rangle_T \geq \delta \|\mathbf{y}\|_{2T}^2 + \varepsilon \|\mathbf{u}\|_{2T}^2$

We need to show: $\exists \bar{\delta}, \bar{\varepsilon} \in \mathbb{R}_{>0}$ such that $\langle \mathbf{y} - D\mathbf{u}, \mathbf{u} \rangle_T \geq \bar{\delta} \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \bar{\varepsilon} \|\mathbf{u}\|_{2T}^2$

Proof:

$$
\begin{aligned}
\langle \mathbf{y}, \mathbf{u} \rangle_T &\geq \delta \|\mathbf{y}\|_{2T}^2 + \varepsilon \|\mathbf{u}\|_{2T}^2 \\
&= \delta \|\mathbf{y}\|_{2T}^2 + \varepsilon \|\mathbf{u}\|_{2T}^2 + 2\delta D \langle \mathbf{y}, \mathbf{u} \rangle_T - 2\delta D \langle \mathbf{y}, \mathbf{u} \rangle_T + \delta D^2 \|\mathbf{u}\|_{2T}^2 - \delta D^2 \|\mathbf{u}\|_{2T}^2 \\
&= \delta \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + 2\delta D \langle \mathbf{y}, \mathbf{u} \rangle_T + \left(\varepsilon - \delta D^2 \right) \|\mathbf{u}\|_{2T}^2 \\
&\geq \delta \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + 2\delta^2 D \|\mathbf{y}\|_{2T}^2 + 2\delta \varepsilon D \|\mathbf{u}\|_{2T}^2 + \left(\varepsilon - \delta D^2 \right) \|\mathbf{u}\|_{2T}^2 \\
&\geq \delta \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \left(\varepsilon + 2\delta \varepsilon D - \delta D^2 \right) \|\mathbf{u}\|_{2T}^2
\end{aligned}
$$

Proof (Contd.)

We need to show: $\exists \bar{\delta}, \bar{\varepsilon} \in \mathbb{R}_{>0}$ such that $\langle \mathbf{y} - D\mathbf{u}, \mathbf{u} \rangle_T \geq \bar{\delta} \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \bar{\varepsilon} \|\mathbf{u}\|_{2T}^2$

$$
\begin{aligned}
\langle \mathbf{y}, \mathbf{u} \rangle_T &\geq \delta \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \left(\varepsilon + 2\delta\varepsilon D - \delta D^2\right) \|\mathbf{u}\|_{2T}^2 \\
&= \delta \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \left(\varepsilon + 2\delta\varepsilon D - \delta D^2\right) \|\mathbf{u}\|_{2T}^2 - D\|\mathbf{u}\|_{2T}^2 + D\|\mathbf{u}\|_{2T}^2 \\
&= \delta \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \left(\varepsilon + (2\delta\varepsilon - 1)D - \delta D^2\right) \|\mathbf{u}\|_{2T}^2 + D\|\mathbf{u}\|_{2T}^2 \\
\langle \mathbf{y}, \mathbf{u} \rangle_T - D\|\mathbf{u}\|_{2T}^2 &\geq \delta \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \left(\varepsilon + (2\delta\varepsilon - 1)D - \delta D^2\right) \|\mathbf{u}\|_{2T}^2 \\
\langle \mathbf{y} - D\mathbf{u}, \mathbf{u} \rangle_T &\geq \underbrace{\delta}_{\bar{\delta}} \|\mathbf{y} - D\mathbf{u}\|_{2T}^2 + \underbrace{\left(\varepsilon + (2\delta\varepsilon - 1)D - \delta D^2\right)}_{\bar{\varepsilon}} \|\mathbf{u}\|_{2T}^2
\end{aligned}
$$

Remark

As $D \to 0$ the VSP property of ∇f is recovered.

Proof: ℓ_2-Stability of Gradient Descent using the Passivity Theorem

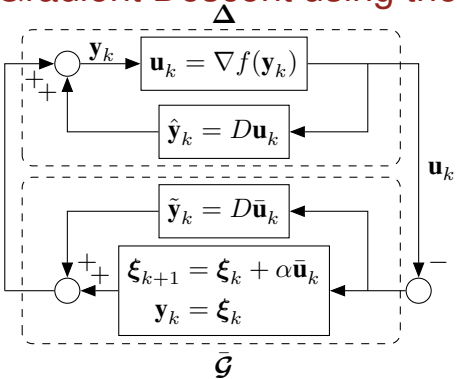


Figure 4: Loop transformation of gradient descent as a negative feedback loop.

- $\bar{\boldsymbol{\mathcal{G}}}$ is passive if and only if $\frac{\alpha}{2} \leq D$.
- Provided $D < \frac{\sqrt{L^4+4L^3m+10L^2m^2+4Lm^3+m^4}-L^2-m^2}{2Lm(L+m)}$, $\boldsymbol{\Delta}$ is VSP.

The ℓ_2-stablility of the negative feedback interconnection follows from the passivity theorem.

Triple Momentum (TM) [Van Scoy, Freeman, Lynch, 2018]

Definition (Triple Momentum (TM))

Consider the update rule

$$\boldsymbol{\xi}_{k+1} = (1+\beta)\boldsymbol{\xi}_k - \beta\boldsymbol{\xi}_{k-1} - \alpha\nabla f(\mathbf{y}_k), \tag{12a}$$
$$\mathbf{y}_k = (1+\gamma)\boldsymbol{\xi}_k - \gamma\boldsymbol{\xi}_{k-1}, \tag{12b}$$
$$\mathbf{x}_k = (1+\delta)\boldsymbol{\xi}_k - \delta\boldsymbol{\xi}_{k-1}. \tag{12c}$$

For $\rho = 1 - 1/\sqrt{\kappa}$, the triple momentum method is defined as the algorithm in (12) with

$$(\alpha, \beta, \gamma, \delta) = \left(\frac{1+\rho}{L}, \frac{\rho^2}{2-\rho}, \frac{\rho^2}{(1+\rho)(2-\rho)}, \frac{\rho^2}{1-\rho^2}\right). \tag{13}$$

Question: What is the minimum D required for Passivity?

Modified TM: $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right] = \left[\begin{array}{cc|c} 1+\beta & -\beta & \alpha \\ 1 & 0 & 0 \\ \hline 1+\gamma & -\gamma & D \end{array}\right]$

Review (Hitz and Anderson, 1969)

$$\mathbf{A}^{\mathsf{T}}\mathbf{P}\mathbf{A} - \mathbf{P} \preceq 0, \tag{4a}$$

$$\mathbf{B}^{\mathsf{T}}\mathbf{P}\mathbf{A} = \mathbf{C}, \tag{4b}$$

$$\mathbf{B}^{\mathsf{T}}\mathbf{P}\mathbf{B} - \left(\mathbf{D}^{\mathsf{T}} + \mathbf{D}\right) \preceq 0. \tag{4c}$$

Remark

The necessary and sufficient conditions for $\left[\begin{array}{cc|c} 1+\beta & -\beta & \alpha \\ 1 & 0 & 0 \\ \hline 1+\gamma & -\gamma & D \end{array}\right]$ *to be passive are:*

1. $\frac{\alpha\gamma}{2\beta} \leq D$.
2. $\beta < \frac{\gamma}{1+\gamma}$.