PERFORMANCE ESTIMATION OF FIRST-ORDER OPTIMIZATION METHODS ON CONVEX FUNCTIONS COMPOSED WITH LINEAR MAPPINGS

Nizar Bousselmi

François Glineur and Julien Hendrickx

Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM)
Université catholique de Louvain (UCLouvain)

Thesis supported by a FRIA grant

Worst-case performance of a method on a class of functions

Common question in optimization :

Worst-case **performance** of an optimization method $\mathcal{M}$ on

$$\min_x f(x)$$

where $f \in \mathcal{F}$ has some properties (smoothness, convexity,...) ?

Common question in optimization :

Worst-case performance of an optimization method $\mathcal{M}$ on

$$\min_x f(x)$$

where $f \in \mathcal{F}$ has some properties (smoothness, convexity,...) ?

Example :

Worst-case performance of $\overbrace{\text{gradient method}}^{\mathcal{M}}$ on $\overbrace{L\text{-smooth convex functions}}^{\mathcal{F}}$ after N iterations ?

$$\overbrace{f(x_N) - f^*}^{\text{performance}} \leq \frac{L}{2} \frac{1}{2N+1}.$$

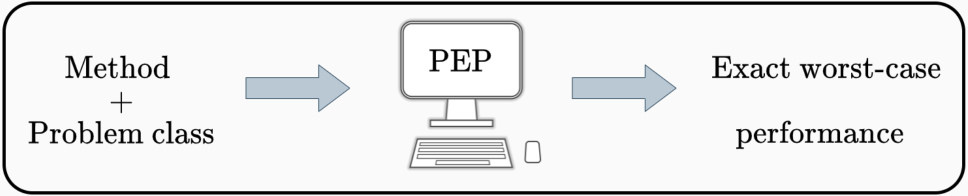
Performance Estimation Problem (PEP)

Theoretical and practical framework to analyze **performance** of optimization methods on problem classes.

- *Performance of first-order methods...*Drori & Teboulle 2013
- *Convex interpolation and performance estimation...*Taylor 2017

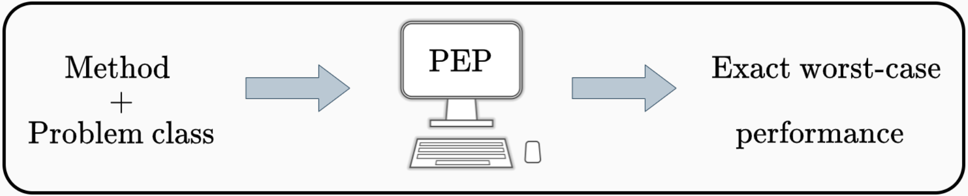
Theoretical and practical framework to analyze **performance** of optimization methods on problem classes.

- *Performance of first-order methods...* Drori & Teboulle 2013
- *Convex interpolation and performance estimation...* Taylor 2017

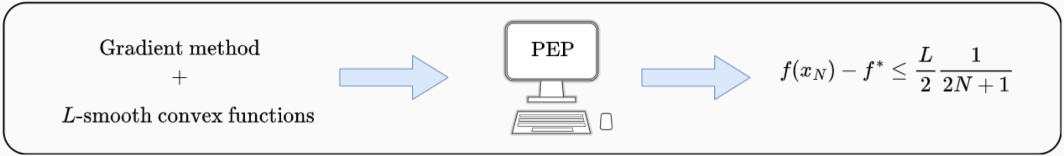


Theoretical and practical framework to analyze **performance** of optimization methods on problem classes.

- *Performance of first-order methods...*Drori & Teboulle 2013
- *Convex interpolation and performance estimation...*Taylor 2017



Example :



PEP IS WELL DEVELOPED

Large number of methods and function classes already analyzed through PEP

- **Smooth convex** and **smooth strongly convex** functions; [Taylor, Hendrickx, Glineur]
- **Constrained** optimization (projected gradient); [Taylor, Glineur, Hendrickx]
- **Non-smooth** optimization (subgradient, proximal operators); [Taylor, Glineur, Hendrickx]
- **Non-convex** and **hypo-convex** functions [Rotaru, Glineur, Patrinos], [Abbaszadehpeivasti, de Klerk, Zamani]
- **Stochastic** optimization;
- **Decentralized** optimization; [Colla, Hendrickx]
- **Coordinate descent** method;
- etc.

PEP IS WELL DEVELOPED

Large number of methods and function classes already analyzed through PEP

- **Smooth convex** and **smooth strongly convex** functions; [Taylor, Hendrickx, Glineur]
- **Constrained** optimization (projected gradient); [Taylor, Glineur, Hendrickx]
- **Non-smooth** optimization (subgradient, proximal operators); [Taylor, Glineur, Hendrickx]
- **Non-convex** and **hypo-convex** functions [Rotaru, Glineur, Patrinos], [Abbaszadehpeivasti, de Klerk, Zamani]
- **Stochastic** optimization;
- **Decentralized** optimization; [Colla, Hendrickx]
- **Coordinate descent** method;
- etc.
- Our contribution: Convex functions composed with **linear mappings**.

Performance Estimation Problem

Interpolation conditions for linear mappings

Exploitation of new tool

Input = optimization method + problem class
Output = worst instance in problem class

Example:

- N steps of gradient method $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$
- L-smooth convex functions f

PEP

$$
\begin{aligned}
\max_{\text{points } x_k, x^*, \text{ function } f} \quad & f(x_N) - f(x^*) \\
\text{s.t.} \quad & f \text{ } L\text{-smooth convex}, \\
& x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \\
& ||x^* - x_0|| \leq 1, \\
& \nabla f(x^*) = 0.
\end{aligned}
$$

Input = optimization method + problem class
Output = worst instance in problem class

Example:

- N steps of gradient method $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$
- L-smooth convex functions f

PEP

$$\max_{\text{points } x_k, x^*, \text{ function } f} \quad f(x_N) - f(x^*)$$

$$\begin{aligned}
\text{s.t.} \quad & f \; L\text{-smooth convex}, \\
& x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \\
& ||x^* - x_0|| \leq 1, \\
& \nabla f(x^*) = 0.
\end{aligned}$$

Input = optimization method + problem class
Output = worst instance in problem class

Example:

- N steps of gradient method $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$
- L-smooth convex functions f

PEP

$$\max_{\text{points } x_k, x^*,\ \text{function } f} \quad f(x_N) - f(x^*)$$

$$\text{s.t.} \quad f \ L\text{-smooth convex},$$

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

$$||x^* - x_0|| \leq 1,$$

$$\nabla f(x^*) = 0.$$

Input = optimization method + problem class
Output = worst instance in problem class

Example:

- N steps of gradient method $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$
- L-smooth convex functions f

PEP

$$
\begin{aligned}
\max_{\text{points } x_k, x^*, \text{ function } f} \quad & f(x_N) - f(x^*) \\
\text{s.t.} \quad & f \ L\text{-smooth convex}, \\
& x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \\
& ||x^* - x_0|| \leq 1, \\
& \nabla f(x^*) = 0.
\end{aligned}
$$

Input = optimization method + problem class
Output = worst instance in problem class

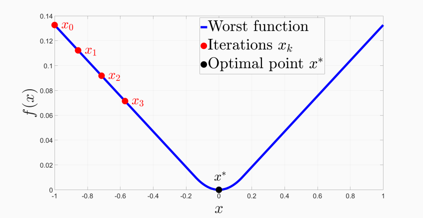
Example:

- N steps of gradient method $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$
- L-smooth convex functions f

PEP solved

Output :

- Worst performance : $f(x_N) - f^* \leq \frac{L}{2}\frac{1}{2N+1}$ for any N;
- Worst function for $N = 3$ and $L = 1$:



PEP as finite-dimensional problem

f **infinite-dimensional** but only access to x_k, $f(x_k)$, $\nabla f(x_k)$ … **black-box** property !

PEP

$$
\begin{aligned}
\max_{\text{points } x_k, x^*, \text{function } f} \quad & f(x_N) - f(x^*) \\
\text{s.t.} \quad & f \ L\text{-smooth convex}, \\
& x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \\
& ||x^* - x_0|| \leq 1, \\
& \nabla f(x^*) = 0.
\end{aligned}
$$

PEP as finite-dimensional problem

f **infinite-dimensional** but only access to x_k, $f(x_k)$, $\nabla f(x_k)$ ... **black-box** property !

PEP

$$\max_{\text{points } x_k, x^*, f_k, f^*, g_k, g^*} \quad f_N - f^*$$

$$\begin{aligned}
\text{s.t.} \quad & \exists f \ L\text{-smooth convex}: \quad f(x_k) = f_k, \quad \nabla f(x_k) = g_k, \\
& \qquad\qquad\qquad\qquad\qquad f(x^*) = f^*, \quad \nabla f(x^*) = g^*, \\
& x_{k+1} = x_k - \frac{1}{L} g_k, \\
& ||x^* - x_0|| \leq 1, \\
& g^* = 0.
\end{aligned}$$

f **infinite-dimensional** but only access to x_k, $f(x_k)$, $\nabla f(x_k)$ ... **black-box** property !

PEP

$$\begin{aligned} \max_{\text{points } x_k, x^*, f_k, f^*, g_k, g^*} \quad & f_N - f^* \\ \text{s.t.} \quad & \exists f \text{ } L\text{-smooth convex}: \quad f(x_k) = f_k, \quad \nabla f(x_k) = g_k, \\ & \qquad\qquad\qquad\qquad\quad f(x^*) = f^*, \quad \nabla f(x^*) = g^*, \\ & x_{k+1} = x_k - \frac{1}{L} g_k, \\ & ||x^* - x_0|| \leq 1, \\ & g^* = 0. \end{aligned}$$

Interpolation condition to reformulate.

PEP as finite-dimensional problem

f **infinite-dimensional** but only access to x_k, $f(x_k)$, $\nabla f(x_k)$ ... **black-box** property !

PEP

$$\max_{\text{points } x_k, x^*, f_k, f^*, g_k, g^*} \quad f_N - f^*$$

$$\text{s.t.} \quad f_j \geq f_k + g_k^T(x_j - x_k) + \frac{1}{2L}||g_j - g_k||^2,$$

$$x_{k+1} = x_k - \frac{1}{L} g_k,$$

$$||x^* - x_0|| \leq 1,$$

$$g^* = 0.$$

Interpolation condition to reformulate.

Can be reformulated as convex **semidefinite problem**, efficiently solvable !

Interpolation conditions for L-smooth convex functions

Given x_k, g_k and f_k $\quad \forall k = 0, \dots, N,$

$\exists$ L-smooth convex f such that $\begin{cases} f(x_k) & = f_k \;\; \forall k = 0, \dots, N, \\ \nabla f(x_k) & = g_k \;\; \forall k = 0, \dots, N, \end{cases}$ if and only if

$$f_j \geq f_k + g_k^T(x_j - x_k) + \frac{1}{2L}||g_j - g_k||^2 \quad \forall j, k = 0, \dots, N.$$

Interpolation conditions for L-smooth convex functions

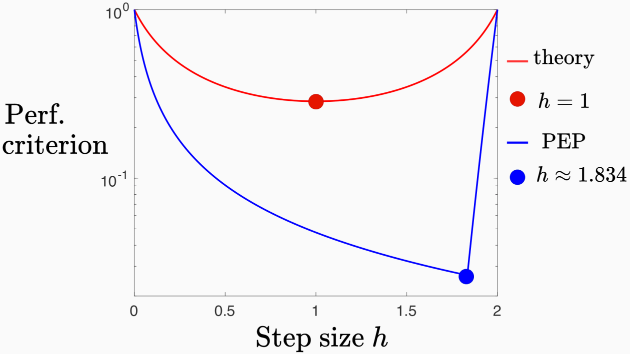
Given x_k, g_k and f_k $\quad \forall k = 0, \ldots, N$,

$\exists$ L-smooth convex f such that $\begin{cases} f(x_k) & = f_k \ \ \forall k = 0, \ldots, N, \\ \nabla f(x_k) & = g_k \ \ \forall k = 0, \ldots, N, \end{cases}$ if and only if

$$f_j \geq f_k + g_k^T(x_j - x_k) + \frac{1}{2L}||g_j - g_k||^2 \quad \forall j, k = 0, \ldots, N.$$

Remark : Interpolation conditions (and PEP formulation) exist for numerous function classes : **non-smooth, L-smooth, convex, μ-strongly convex,etc**

Accuracy after 10 steps of gradient method on L-smooth convex functions for varying step size $\frac{h}{L}$



Remarks :

- Theory [Nesterov98] suggests a step size of $\frac{1}{L}$ while PEP recommends $\approx \frac{1.834}{L}$ to optimize the worst-case;
- PEP provides tight results;
- PEP guarantees a performance $\approx$ 10 times better than theory.

Outline

Performance Estimation Problem

Interpolation conditions for linear mappings

Exploitation of new tool

Goal : Analyze worst performance of methods on $\min_x f(x)$ for f using linear mapping.

Goal : Analyze worst performance of methods on $\min_x f(x)$ for f using linear mapping.

Function class	Hypothesis
$f(x) = g(Ax)$	g smooth (strongly) convex
$f(x) = g(Ax) = \frac{1}{2}x^T QX$ (quadratic functions)	$g(y) = \frac{1}{2}\|\|y\|\|^2$ and $Q = A^T A$
$f(x) = h(x) + g(Ax)$	h, g smooth (strongly) convex

Goal : Analyze worst performance of methods on $\min_x f(x)$ for f using linear mapping.

Function class	Hypothesis
$f(x) = g(Ax)$	g smooth (strongly) convex
$f(x) = g(Ax) = \frac{1}{2}x^T QX$ (quadratic functions)	$g(y) = \frac{1}{2}\|\|y\|\|^2$ and $Q = A^T A$
$f(x) = h(x) + g(Ax)$	h, g smooth (strongly) convex

Goal : Analyze worst performance of methods on $\min_x f(x)$ for f using linear mapping.

Function class	Hypothesis
$f(x) = g(Ax)$	g smooth (strongly) convex
$f(x) = g(Ax) = \frac{1}{2}x^T QX$ (quadratic functions)	$g(y) = \frac{1}{2}\|\|y\|\|^2$ and $Q = A^T A$
$f(x) = h(x) + g(Ax)$	h, g smooth (strongly) convex

Goal : Analyze worst performance of methods on $\min_x f(x)$ for f using linear mapping.

Function class	Hypothesis
$f(x) = g(Ax)$	g smooth (strongly) convex

Gradient method on $f(x) = g(Ax)$ with $A = A^T$: $x_{k+1} = x_k - \frac{h}{L} A \nabla g(Ax_k)$

Goal : Analyze worst performance of methods on $\min_x f(x)$ for f using linear mapping.

Function class	Hypothesis
$f(x) = g(Ax)$	g smooth (strongly) convex

Gradient method on $f(x) = g(Ax)$ with $A = A^T$: $x_{k+1} = x_k - \frac{h}{L} A \nabla g(Ax_k)$

or equivalently, decomposing operations for each step

$$
\begin{cases}
y_k & = Ax_k \\
u_k & = \nabla g(y_k) \\
v_k & = Au_k \\
x_{k+1} & = x_k - \frac{h}{L} v_k
\end{cases}
$$

Goal : Analyze worst performance of methods on $\min_x f(x)$ for f using linear mapping.

Function class	Hypothesis
$f(x) = g(Ax)$	g smooth (strongly) convex

Gradient method on $f(x) = g(Ax)$ with $A = A^T$: $\quad x_{k+1} = x_k - \frac{h}{L} A \nabla g(Ax_k)$

or equivalently, decomposing operations for each step

$$\begin{cases} y_k &= Ax_k \quad \text{New interpolation conditions} \\ u_k &= \nabla g(y_k) \quad \text{Standard interpolation conditions} \\ v_k &= Au_k \quad \text{New interpolation conditions} \\ x_{k+1} &= x_k - \frac{h}{L} v_k \quad \text{Standard} \end{cases}$$

Let $N \in \mathbb{N}$, $S = \{0, \dots, N\}$, $X = (x_0 \ \cdots \ x_N)$, $Y = (y_0 \ \cdots \ y_N)$ and $L \in \mathbb{R}$.

Theorem (Symmetric matrix with spectrum between 0 and L)

Given x_k and y_k $\ \forall k \in S$ and $G = (X\ Y)^T (X\ Y) = \begin{pmatrix} X^T X & X^T Y \\ Y^T X & Y^T Y \end{pmatrix} \triangleq \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succeq 0.$

If $\exists Q$ symmetric $\ : \ 0 \preceq Q \preceq L$ and $y_k = Q x_k \ \forall k = 0, \dots, N$ then

$$\begin{cases} B = B^T, \\ B \succeq \frac{C}{L}. \end{cases}$$

Let $N \in \mathbb{N}$, $S = \{0, \ldots, N\}$, $X = (x_0 \ \cdots \ x_N)$, $Y = (y_0 \ \cdots \ y_N)$ and $\mu \leq L \in \mathbb{R}$.

Theorem (Symmetric matrix with spectrum between μ and L)

Given x_k and y_k $\forall k \in S$ and $G = (X\ Y)^T(X\ Y) = \begin{pmatrix} X^TX & X^TY \\ Y^TX & Y^TY \end{pmatrix} \triangleq \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succeq 0.$

If $\exists Q$ symmetric $: \mu I \preceq Q \preceq LI$ *and* $y_k = Qx_k \ \ \forall k = 0, \ldots, N$ *then*

$$\begin{cases} B = B^T, \\ B \succeq \frac{\mu L}{\mu+L} A + \frac{1}{\mu+L} C. \end{cases}$$

Let $N \in \mathbb{N}$, $S = \{0, \dots, N\}$, $X = (x_0 \;\cdots\; x_N)$, $Y = (y_0 \;\cdots\; y_N)$ and $\mu \leq L \in \mathbb{R}$.

Theorem (Symmetric matrix with spectrum between μ and L)

Given x_k and y_k $\forall k \in S$ and $G = (X\ Y)^T(X\ Y) = \begin{pmatrix} X^TX & X^TY \\ Y^TX & Y^TY \end{pmatrix} \triangleq \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succeq 0.$

If $\exists Q$ *symmetric* $: \mu I \preceq Q \preceq LI$ *and* $y_k = Qx_k \;\; \forall k = 0, \dots, N$ *then*

$$\begin{cases} B = B^T, \\ B \succeq \frac{\mu L}{\mu+L} A + \frac{1}{\mu+L} C. \end{cases}$$

Remark :

- We can have the «**only if**» with a slightly modified theorem;
- We only consider homogeneous quadratic functions;
- A similar theorem exists for **non-symmetric** matrix with bounded singular values.

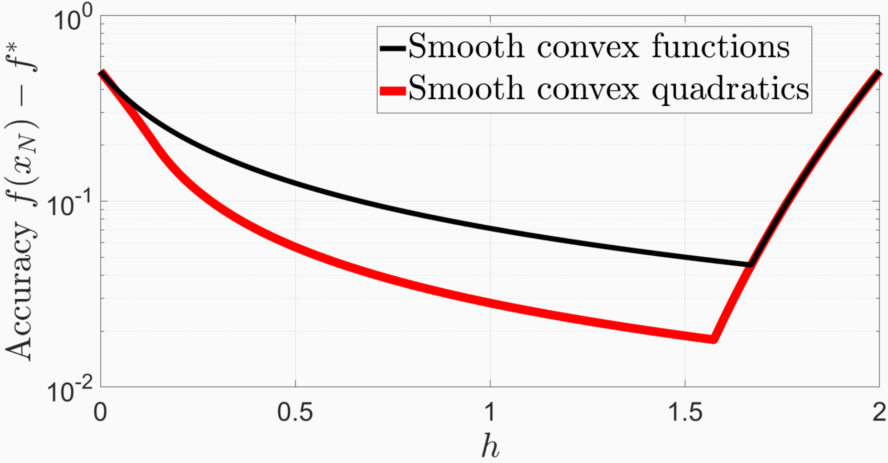
Performance Estimation Problem

Interpolation conditions for linear mappings

Exploitation of new tool

Worst performance of $N = 3$ steps of gradient method on $\min_x f(x)$ w.r.t. the step size h

- $f(x) = g(Ax) = \frac{1}{2}x^TQx$, $g(y) = \frac{1}{2}||y||^2$
- $Q = A^TA$ with $0 \preceq Q \preceq I$



Smooth convex functions	Smooth convex quadratic
$f(x_N) - f^* \leq \frac{LR^2}{2} \begin{cases} \frac{1}{2Nh+1} & \text{if } h \in [0, h_0] \\ (1-h)^{2N} & \text{if } h \in [h_0, \infty] \end{cases}$	$f(x_N) - f^* \leq \frac{LR^2}{2} \begin{cases} (1-h)^{2N} & \text{if } h \in [0, \frac{1}{2N+1}] \\ \frac{1}{h}\frac{(2N)^{2N}}{(2N+1)^{2N+1}} & \text{if } h \in [\frac{1}{2N+1}, h_1] \\ (1-h)^{2N} & \text{if } h \in [h_1, \infty] \end{cases}$

Note : As in the general case, worst quadratic functions are one-dimensional.

Independent existing approach to analyze first-order methods on quadratic functions

Worst-case performance of a first-order method on quadratic functions is

Polynomial approach

$$\max_{\rho\in[\mu,L]} \frac{\rho}{2} R\big(1+\rho K_N(\rho)\big)^2$$

where K_N is a polynomial of degree N that depends explicitly on the method.

Independent existing approach to analyze first-order methods on quadratic functions

Worst-case performance of a first-order method on quadratic functions is

Polynomial approach

$$\max_{\rho\in[\mu,L]} \frac{\rho}{2} R\left(1+\rho K_N(\rho)\right)^2$$

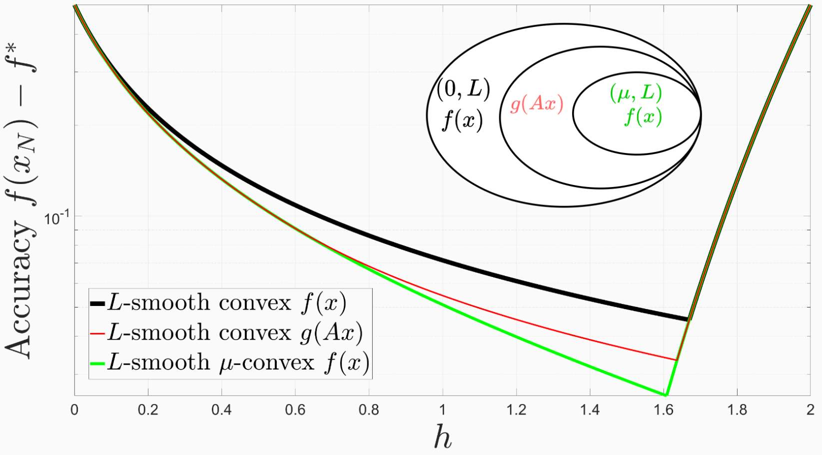
where K_N is a polynomial of degree N that depends explicitly on the method.

It only works for quadratic functions !

Function class	PEP	Polynomial approach
Quadratics $\frac{1}{2}x^T Q x$	OK	OK
g(Ax)	OK	KO
f(x) + g(Ax)	OK	KO
Class with matrices (except quadratics)	OK	KO

Worst performance of $N = 3$ steps of gradient method on $\min_x f(x)$ w.r.t. the step size h

- $f(x) = \mathbf{g(Ax)}$, $g(y)$ smooth strongly convex
- $0 \preceq A \preceq I$



Smooth convex $f(x)$	Smooth convex $g(Ax)$
$\frac{LR^2}{2}\begin{cases}\frac{1}{2Nh+1} & \text{if } h\in[0,h_0]\\ (1-h)^{2N} & \text{if } h\in[h_0,\infty]\end{cases}$	$\frac{LR^2}{2}\begin{cases}\frac{\kappa_g}{\kappa_g-1+(1-\kappa_g h)^{-2N}} & \text{if } h\in[0,h_2]\\ \frac{\kappa_g\frac{h_0}{h}}{\kappa_g-1+(1-\kappa_g h_0)^{-2N}} & \text{if } h\in[h_2,h_3]\\ (1-h)^{2N} & \text{if } h\in[h_3,\infty]\end{cases}$

State of the art :

- *PEP* gives **worst-case performance** of **methods** on a **function class** (for which interpolation conditions are available).

Our contribution :

- Extending PEP to methods and classes using linear mapping;
- Analyzing $\frac{1}{2}x^T QX$, $g(Ax)$.

Future research : Analyzing more complex problems and methods (e.g. $f(x) + g(Ax)$, Chambolle-Pock, Condat-Vu) and identifying performance.

Definitions and notations

f is L-smooth when

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||.$$

First-order method of the form

$$x_N = x_0 - \sum_{i=0}^{N-1} h_{N,i} \nabla f(x_i).$$

Case $\mu = L$

Let $G = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succeq 0$ and $\mu = L \in \mathbb{R}$.

Theorem

G can be written as $\begin{pmatrix} X^TX & X^TQX \\ X^TQX & X^TQ^2X \end{pmatrix}$ for a symmetric matrix Q with $LI \preceq Q \preceq LI$ if and only if

$$B = B^T,$$
$$C \preceq L^2 A.$$

Interpolation condition for L-smooth convex functions

f L-smooth convex if and only if

$$f(x) \geq f(y) + \nabla f^T(y)(x-y) + \frac{1}{2L}||\nabla f(x) - \nabla f(y)||^2 \quad \forall x, y$$

f L-smooth convex : $f(x_k) = f_k, \; \nabla f(x_k) = g_k$ if and only if

$$f_i \geq f_j + g_j^T(x_i - x_j) + \frac{1}{2L}||g_i - g_j||^2 \quad \forall i, j$$

SDP formulation

N steps of gradient method on L-smooth convex functions.

Matrix variable: $G = (g_0 \ \dots \ g_N \ x_0)^T (g_0 \ \dots \ g_N \ x_0) \in \mathbb{S}^{N+2}$

Parameters:

- $h_i = (0 \ \dots 0 \ \frac{-1}{L} \ 0 \ \dots \ 0 \ 1) \in \mathbb{R}^{N+2}$
- $u_i = (0 \dots 0 \ 1 \ 0 \ \dots 0) \in \mathbb{R}^{N+2}$
- $2A_{ij} = u_j(h_i - h_j)^T + (h_i - h_j)u_j^T + \frac{1}{L}(u_i - u_j)(u_i - u_j)^T$
- $A_R = u_{N+1}u_{N+1}^T$

$$
\begin{aligned}
\max_{G \in \mathbb{S}^{N+2}, f \in \mathbb{R}^{N+1}} \quad & f_N - f^* \\
\text{s.t.} \quad & f_j - f_i + \mathrm{Tr}(GA_{ij}) \leq 0, \quad \forall i, j \\
& \mathrm{Tr}(GA_{ij}) - R^2 \leq 0, \quad \forall i, j \\
& G \succeq 0.
\end{aligned}
$$