Computer-aided analyses of optimization methods via potential functions

Adrien Taylor

CWI-Inria workshop - September 2020

4TUNE

Adaptive, Efficient, Provable and Flexible Tuning for Machine Learning

Joint research team between CWI and Inria.

Team

4TUNE includes 2 research scientists from the Centrum Wiskunde & Informatica (CWI) and 3 researchers from the Sierra project-team of Inria.

CWI researchers


Peter Grünwald


Wouter M. Koolen

INRIA, Sierra project-team researchers


Francis Bach


Pierre Gaillard


Adrien Taylor

Newborn
in the CWI-Inria lab!

Long-term goal: push adaptive machine learning to the next level.

We aim to develop refined methods, going beyond traditional worst-case analysis, for exploiting structure in the learning problem at hand [...]

Francis Bach

“Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions” (COLT 2019).

What is this presentation about?

What is this presentation about?

Computer-assisted analyses of first-order optimization methods

What is this presentation about?

Computer-assisted analyses of first-order optimization methods

(Drori & Teboulle 2014), (Lessard, Recht & Packard 2016), (T, Hendrickx & Glineur 2017),

and few others.

What is this presentation about?

Computer-assisted analyses of first-order optimization methods

(Drori & Teboulle 2014), (Lessard, Recht & Packard 2016), (T, Hendrickx & Glineur 2017), and few others.

Focus on *simple* proofs, relying on (quadratic) *potential functions*

(Nesterov 1983), (Beck & Teboulle 2009), (Wilson, Recht & Jordan 2016), (Hu & Lessard 2017), (Bansal & Gupta 2019), and many others.

Example: analysis of a gradient step

Find $x_\star$ such that

$$f(x_\star) = \min_x f(x).$$

Example: analysis of a gradient step

Find $x_\star$ such that

$$f(x_\star) = \min_x f(x).$$

(Gradient method) We decide to use: $x_{k+1} = x_k - \gamma_k f'(x_k)$

Example: analysis of a gradient step

Find $x_\star$ such that

$$f(x_\star) = \min_x f(x).$$

(Gradient method) We decide to use: $x_{k+1} = x_k - \gamma_k f'(x_k)$

Question: what *a priori* guarantees after N iterations?

Example: analysis of a gradient step

Find $x_\star$ such that

$$f(x_\star) = \min_x f(x).$$

(Gradient method) We decide to use: $x_{k+1} = x_k - \gamma_k f'(x_k)$

Question: what *a priori* guarantees after N iterations?

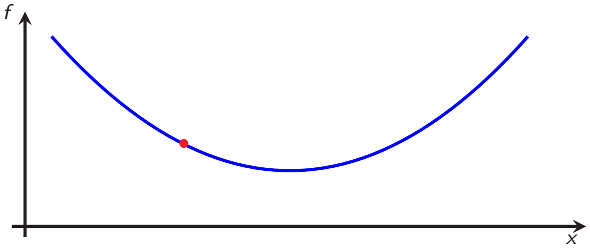
Examples: what about $f(x_N) - f(x_\star)$, $\|f'(x_N)\|$, $\|x_N - x_\star\|$?

Smooth convex functions

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f is convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:

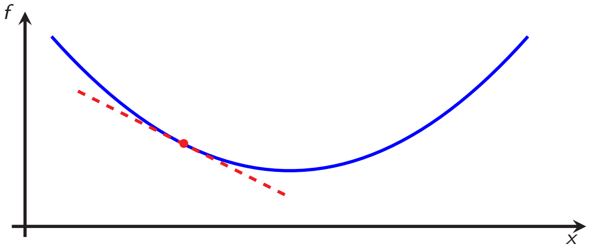
Smooth convex functions

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f is convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:



Smooth convex functions

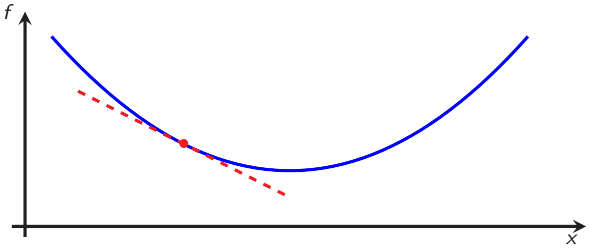
A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f is convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:



(1) (Convexity) $f(x) \geqslant f(y) + \langle f'(y), x - y \rangle$,

Smooth convex functions

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f is convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:

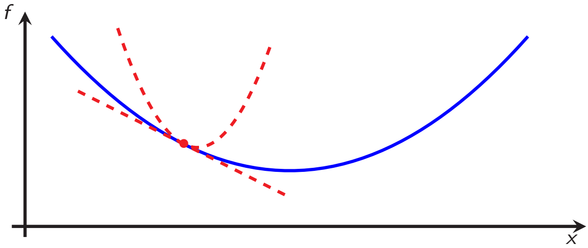


(1) (Convexity) $f(x) \geqslant f(y) + \langle f'(y), x - y \rangle$,

(2) (L-smoothness) $\|f'(x) - f'(y)\| \leqslant L\|x - y\|$,

Smooth convex functions

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f is convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:



(1) (Convexity) $f(x) \geqslant f(y) + \langle f'(y), x - y \rangle$,

(2) (L-smoothness) $\|f'(x) - f'(y)\| \leqslant L\|x - y\|$,

(2b) (L-smoothness) $f(x) \leqslant f(y) + \langle f'(y), x - y \rangle + \frac{L}{2}\|x - y\|^2$.

Potential functions

What guarantees for gradient descent when minimizing a L-smooth convex function

$$f_\star = \min_x f(x)?$$

Potential functions

What guarantees for gradient descent when minimizing a L-smooth convex function

$$f_\star = \min_x f(x)?$$

It is known that $f(x_N) - f_\star = O(\frac{1}{N})$ with small enough step sizes (e.g., $\frac{1}{L}$).

Potential functions

What guarantees for gradient descent when minimizing a L-smooth convex function

$$f_\star = \min_x f(x)?$$

It is known that $f(x_N) - f_\star = O(\frac{1}{N})$ with small enough step sizes (e.g., $\frac{1}{L}$).

For all L-smooth convex f, iterate x_k, and $k \geq 0$, easy to show $\phi^f_{k+1} \leq \phi^f_k$ with

$$\phi^f_k = k(f(x_k) - f_\star) + \frac{L}{2}\|x_k - x_\star\|^2 \textit{ (potential at iteration } k\textit{)},$$

see e.g., (Bansal & Gupta 2019).

Potential functions

What guarantees for gradient descent when minimizing a L-smooth convex function

$$f_\star = \min_x f(x)?$$

It is known that $f(x_N) - f_\star = O(\frac{1}{N})$ with small enough step sizes (e.g., $\frac{1}{L}$).

For all L-smooth convex f, iterate x_k, and $k \geq 0$, easy to show $\phi_{k+1}^f \leq \phi_k^f$ with

$$\phi_k^f = k(f(x_k) - f_\star) + \frac{L}{2}\|x_k - x_\star\|^2 \textit{ (potential at iteration } k\textit{)},$$

see e.g., (Bansal & Gupta 2019).

Why is that nice? Very simple resulting proof:

Potential functions

What guarantees for gradient descent when minimizing a L-smooth convex function

$$f_\star = \min_x f(x)?$$

It is known that $f(x_N) - f_\star = O(\frac{1}{N})$ with small enough step sizes (e.g., $\frac{1}{L}$).

For all L-smooth convex f, iterate x_k, and $k \geq 0$, easy to show $\phi^f_{k+1} \leq \phi^f_k$ with

$$\phi^f_k = k(f(x_k) - f_\star) + \frac{L}{2}\|x_k - x_\star\|^2 \textit{ (potential at iteration } k\textit{)},$$

see e.g., (Bansal & Gupta 2019).

Why is that nice? Very simple resulting proof:

$$\phi^f_N \leq \phi^f_{N-1} \leq \ldots \leq \phi^f_0$$

Potential functions

What guarantees for gradient descent when minimizing a L-smooth convex function

$$f_\star = \min_x f(x)?$$

It is known that $f(x_N) - f_\star = O(\frac{1}{N})$ with small enough step sizes (e.g., $\frac{1}{L}$).

For all L-smooth convex f, iterate x_k, and $k \geq 0$, easy to show $\phi^f_{k+1} \leq \phi^f_k$ with

$$\phi^f_k = k(f(x_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 \textit{ (potential at iteration } k\textit{)},$$

see e.g., (Bansal & Gupta 2019).

Why is that nice? Very simple resulting proof:

$$N(f(x_N) - f_\star) \leq \phi^f_N \leq \phi^f_{N-1} \leq \ldots \leq \phi^f_0$$

Potential functions

What guarantees for gradient descent when minimizing a L-smooth convex function

$$f_\star = \min_x f(x)?$$

It is known that $f(x_N) - f_\star = O(\frac{1}{N})$ with small enough step sizes (e.g., $\frac{1}{L}$).

For all L-smooth convex f, iterate x_k, and $k \geq 0$, easy to show $\phi^f_{k+1} \leq \phi^f_k$ with

$$\phi^f_k = k(f(x_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 \textit{ (potential at iteration } k\textit{)},$$

see e.g., (Bansal & Gupta 2019).

Why is that nice? Very simple resulting proof:

$$N(f(x_N) - f_\star) \leq \phi^f_N \leq \phi^f_{N-1} \leq \ldots \leq \phi^f_0 = \tfrac{L}{2}\|x_0 - x_\star\|^2,$$

Potential functions

What guarantees for gradient descent when minimizing a L-smooth convex function

$$f_\star = \min_x f(x)?$$

It is known that $f(x_N) - f_\star = O(\frac{1}{N})$ with small enough step sizes (e.g., $\frac{1}{L}$).

For all L-smooth convex f, iterate x_k, and $k \geq 0$, easy to show $\phi^f_{k+1} \leq \phi^f_k$ with

$$\phi^f_k = k(f(x_k) - f_\star) + \frac{L}{2}\|x_k - x_\star\|^2 \textit{ (potential at iteration } k\textit{)},$$

see e.g., (Bansal & Gupta 2019).

Why is that nice? Very simple resulting proof:

$$N(f(x_N) - f_\star) \leq \phi^f_N \leq \phi^f_{N-1} \leq \ldots \leq \phi^f_0 = \frac{L}{2}\|x_0 - x_\star\|^2,$$

hence: $f(x_N) - f_\star \leq \frac{L\|x_0 - x_\star\|^2}{2N}$.

How does it work for the gradient method?

Gradient descent, take II: how to bound $\|f'(x_N)\|^2$ using potentials?

How does it work for the gradient method?

Gradient descent, take II: how to bound $\|f'(x_N)\|^2$ using potentials?

Key idea: forget how x_k was generated and prove $\phi^f_{k+1} \leq \phi^f_k$.

- ☺ only need to study one iteration
- ☹ where does this ϕ^f_k comes from!? (structure and dependence on k)

How does it work for the gradient method?

Gradient descent, take II: how to bound $\|f'(x_N)\|^2$ using potentials?

Key idea: forget how x_k was generated and prove $\phi_{k+1}^f \leq \phi_k^f$.

- ☺ only need to study one iteration
- ☹ where does this ϕ_k^f comes from!? (structure and dependence on k)

Starting point: candidate quadratic ϕ_k^f with *all the available information* at iteration k

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k \left(f(x_k) - f_\star\right).$$

How does it work for the gradient method?

Gradient descent, take II: how to bound $\|f'(x_N)\|^2$ using potentials?

Key idea: forget how x_k was generated and prove $\phi_{k+1}^f \leq \phi_k^f$.

- ☺ only need to study one iteration
- ☹ where does this ϕ_k^f comes from!? (structure and dependence on k)

Starting point: candidate quadratic ϕ_k^f with *all the available information* at iteration k

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star \rangle + d_k \left(f(x_k) - f_\star\right).$$

How to choose a_k, b_k, c_k, d_k's?

How does it work for the gradient method?

Gradient descent, take II: how to bound $\|f'(x_N)\|^2$ using potentials?

Key idea: forget how x_k was generated and prove $\phi_{k+1}^f \leq \phi_k^f$.

- ☺ only need to study one iteration
- ☹ where does this ϕ_k^f comes from!? (structure and dependence on k)

Starting point: candidate quadratic ϕ_k^f with *all the available information* at iteration k

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k (f(x_k) - f_\star).$$

How to choose a_k, b_k, c_k, d_k's?

1. choice should satisfy "$\phi_{k+1}^f \leq \phi_k^f$",

How does it work for the gradient method?

Gradient descent, take II: how to bound $\|f'(x_N)\|^2$ using potentials?

Key idea: forget how x_k was generated and prove $\phi_{k+1}^f \leq \phi_k^f$.

- ☺ only need to study one iteration
- ☹ where does this ϕ_k^f comes from!? (structure and dependence on k)

Starting point: candidate quadratic ϕ_k^f with *all the available information* at iteration k

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k \left(f(x_k) - f_\star\right).$$

How to choose a_k, b_k, c_k, d_k's?

1. choice should satisfy "$\phi_{k+1}^f \leq \phi_k^f$",
2. choice should result in bound on $\|f'(x_N)\|^2$.

How does it work for the gradient method?

Given ϕ^f_{k+1}, ϕ^f_k, *how to verify* that for all L-smooth convex f and iterate x_k

$$\phi^f_{k+1} \leq \phi^f_k?$$

How does it work for the gradient method?

Given ϕ^f_{k+1}, ϕ^f_k, *how to verify* that for all L-smooth convex f and iterate x_k

$$\phi^f_{k+1} \leq \phi^f_k?$$

Notations: the set of such pairs (ϕ^f_k, ϕ^f_{k+1}) is denoted $\mathcal{V}_k$ (here: for gradient method).

How does it work for the gradient method?

Given ϕ_{k+1}^f, ϕ_k^f, *how to verify* that for all L-smooth convex f and iterate x_k

$$\phi_{k+1}^f \leq \phi_k^f?$$

Notations: the set of such pairs (ϕ_k^f, ϕ_{k+1}^f) is denoted $\mathcal{V}_k$ (here: for gradient method).

Answer:

$$\phi_{k+1}^f \leq \phi_k^f \text{ for all } L\text{-smooth convex } f\text{, and iterate } x_k$$

$$\Leftrightarrow$$

some small-sized *linear matrix inequality (LMI)* is feasible.

How does it work for the gradient method?

Given ϕ^f_{k+1}, ϕ^f_k, *how to verify* that for all L-smooth convex f and iterate x_k

$$\phi^f_{k+1} \leq \phi^f_k?$$

Notations: the set of such pairs (ϕ^f_k, ϕ^f_{k+1}) is denoted $\mathcal{V}_k$ (here: for gradient method).

Answer:

$$\phi^f_{k+1} \leq \phi^f_k \text{ for all } L\text{-smooth convex } f\text{, and iterate } x_k$$

$$\Leftrightarrow$$

some small-sized *linear matrix inequality (LMI)* is feasible.

Furthermore: LMI is linear in parameters $\{a_k, b_k, c_k, d_k\}_k$.

How does it work for the gradient method?

Given ϕ_{k+1}^f, ϕ_k^f, *how to verify* that for all L-smooth convex f and iterate x_k

$$\phi_{k+1}^f \leq \phi_k^f?$$

Notations: the set of such pairs (ϕ_k^f, ϕ_{k+1}^f) is denoted $\mathcal{V}_k$ (here: for gradient method).

Answer:

$$\phi_{k+1}^f \leq \phi_k^f \text{ for all } L\text{-smooth convex } f, \text{ and iterate } x_k$$

$$\Leftrightarrow$$

some small-sized *linear matrix inequality (LMI)* is feasible.

Furthermore: LMI is linear in parameters $\{a_k, b_k, c_k, d_k\}_k$.

In others words: *efficient (convex) representation of $\mathcal{V}_k$ available*!

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star \rangle + d_k (f(x_k) - f_\star).$$

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k \left(f(x_k) - f_\star\right).$$

with $\phi_0^f = L^2\|x_0 - x_\star\|^2$ and $\phi_N^f = b_N \|f'(x_N)\|^2$.

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star \rangle + d_k \left(f(x_k) - f_\star\right).$$

with $\phi_0^f = L^2\|x_0 - x_\star\|^2$ and $\phi_N^f = b_N \|f'(x_N)\|^2$.

Motivation: this structure would result in $\|f'(x_N)\|^2 \leq \frac{L^2\|x_0 - x_\star\|^2}{b_N}$.

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k (f(x_k) - f_\star).$$

with $\phi_0^f = L^2\|x_0 - x_\star\|^2$ and $\phi_N^f = b_N \|f'(x_N)\|^2$.

Motivation: this structure would result in $\|f'(x_N)\|^2 \leq \frac{L^2\|x_0-x_\star\|^2}{b_N}$.

Question: largest provable b_N using such potentials?

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k (f(x_k) - f_\star).$$

with $\phi_0^f = L^2\|x_0 - x_\star\|^2$ and $\phi_N^f = b_N \|f'(x_N)\|^2$.

Motivation: this structure would result in $\|f'(x_N)\|^2 \leq \frac{L^2\|x_0 - x_\star\|^2}{b_N}$.

Question: largest provable b_N using such potentials?

$$\max_{\phi_1^f,\ldots,\phi_{N-1}^f, b_N} b_N \text{ such that } (\phi_0^f, \phi_1^f) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \mathcal{V}_{N-1}$$

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k\,(f(x_k) - f_\star).$$

with $\phi_0^f = L^2\|x_0 - x_\star\|^2$ and $\phi_N^f = b_N \|f'(x_N)\|^2$.

Motivation: this structure would result in $\|f'(x_N)\|^2 \leq \frac{L^2\|x_0 - x_\star\|^2}{b_N}$.

Question: largest provable b_N using such potentials?

$$\max_{\phi_1^f,\ldots,\phi_{N-1}^f, b_N} b_N \text{ such that } (\phi_0^f, \phi_1^f) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \mathcal{V}_{N-1}$$

Let's engineer a worst-case guarantee:

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k (f(x_k) - f_\star).$$

with $\phi_0^f = L^2\|x_0 - x_\star\|^2$ and $\phi_N^f = b_N \|f'(x_N)\|^2$.

Motivation: this structure would result in $\|f'(x_N)\|^2 \leq \frac{L^2\|x_0 - x_\star\|^2}{b_N}$.

Question: largest provable b_N using such potentials?

$$\max_{\phi_1^f,\ldots,\phi_{N-1}^f, b_N} b_N \text{ such that } (\phi_0^f, \phi_1^f) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \mathcal{V}_{N-1}$$

Let's engineer a worst-case guarantee:

1. Solve the SDP for some values of N.

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k \left(f(x_k) - f_\star\right).$$

with $\phi_0^f = L^2\|x_0 - x_\star\|^2$ and $\phi_N^f = b_N \|f'(x_N)\|^2$.

Motivation: this structure would result in $\|f'(x_N)\|^2 \leq \frac{L^2\|x_0 - x_\star\|^2}{b_N}$.

Question: largest provable b_N using such potentials?

$$\max_{\phi_1^f,\ldots,\phi_{N-1}^f, b_N} b_N \text{ such that } (\phi_0^f, \phi_1^f) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \mathcal{V}_{N-1}$$

Let's engineer a worst-case guarantee:

1. Solve the SDP for some values of N.
2. Observe the a_k, b_k, c_k, d_k's for some values of N.

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k (f(x_k) - f_\star).$$

with $\phi_0^f = L^2\|x_0 - x_\star\|^2$ and $\phi_N^f = b_N \|f'(x_N)\|^2$.

Motivation: this structure would result in $\|f'(x_N)\|^2 \leq \frac{L^2\|x_0 - x_\star\|^2}{b_N}$.

Question: largest provable b_N using such potentials?

$$\max_{\phi_1^f,\ldots,\phi_{N-1}^f, b_N} b_N \text{ such that } (\phi_0^f, \phi_1^f) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \mathcal{V}_{N-1}$$

Let's engineer a worst-case guarantee:

1. Solve the SDP for some values of N.
2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.

How does it work for the gradient method?

Recap: we want to bound $\|f'(x_N)\|^2$; choose

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k \left(f(x_k) - f_\star\right).$$

with $\phi_0^f = L^2\|x_0 - x_\star\|^2$ and $\phi_N^f = b_N \|f'(x_N)\|^2$.

Motivation: this structure would result in $\|f'(x_N)\|^2 \leq \frac{L^2\|x_0-x_\star\|^2}{b_N}$.

Question: largest provable b_N using such potentials?

$$\max_{\phi_1^f,\ldots,\phi_{N-1}^f,b_N} b_N \text{ such that } (\phi_0^f,\phi_1^f)\in\mathcal{V}_0,\ldots,(\phi_{N-1}^f,\phi_N^f)\in\mathcal{V}_{N-1}$$

Let's engineer a worst-case guarantee:

1. Solve the SDP for some values of N.
2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.
4. Prove target result by analytically playing with $\mathcal{V}_k$ (i.e., study single iteration).

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2 \|x_0 - x_\star\|^2}{b_N}.$$

Numerically (live if time allows)

$$\begin{aligned} N &= \\ b_N &= \end{aligned}$$

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\,\|x_0 - x_\star\|^2}{b_N}.$$

Numerically (live if time allows)

$$\begin{aligned} N &= \quad 1 \\ b_N &= \end{aligned}$$

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\,\|x_0 - x_\star\|^2}{b_N}.$$

Numerically (live if time allows)

$$\begin{aligned} N &= \quad 1 \\ b_N &= \quad 4 \end{aligned}$$

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\,\|x_0 - x_\star\|^2}{b_N}.$$

   Numerically (live if time allows)

$$\begin{array}{rcc} N = & 1 & 2 \\ b_N = & 4 & 9 \end{array}$$

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\,\|x_0 - x_\star\|^2}{b_N}.$$

Numerically (live if time allows)

$$\begin{array}{rccc} N = & 1 & 2 & 3 \\ b_N = & 4 & 9 & 16 \end{array}$$

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\,\|x_0 - x_\star\|^2}{b_N}.$$

Numerically (live if time allows)

$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \ldots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \ldots & 10201 \end{array}$$

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\,\|x_0 - x_\star\|^2}{b_N}.$$

   Numerically (live if time allows)

$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \ldots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \ldots & 10201 \end{array}$$

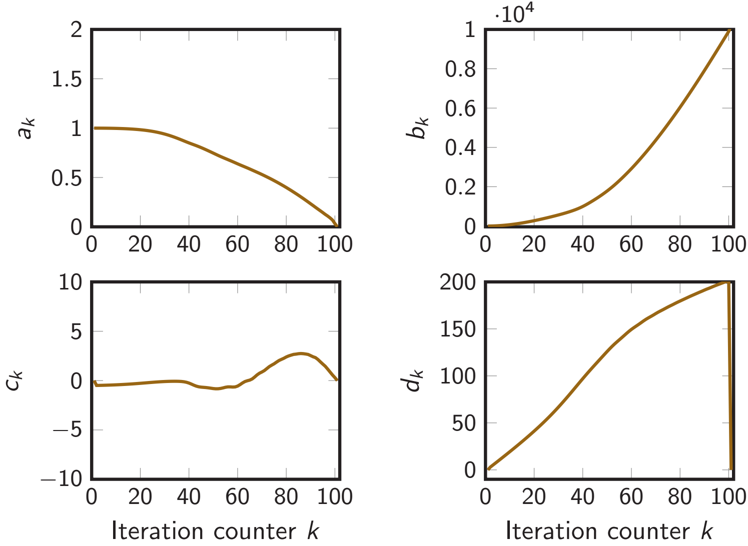
2. Observe the a_k, b_k, c_k, d_k's for some values of N.

Fixed horizon $N = 100$, $L = 1$, and

$$\phi_k^f = a_k \left\|x_k - x_\star\right\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \left\langle f'(x_k), x_k - x_\star\right\rangle + d_k \left(f(x_k) - f_\star\right).$$

Fixed horizon $N = 100$, $L = 1$, and

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \|f'(x_k)\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k (f(x_k) - f_\star).$$



How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2 \left\|x_0 - x_\star\right\|^2}{b_N}.$$

   Numerically (live if time allows)

$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \ldots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \ldots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

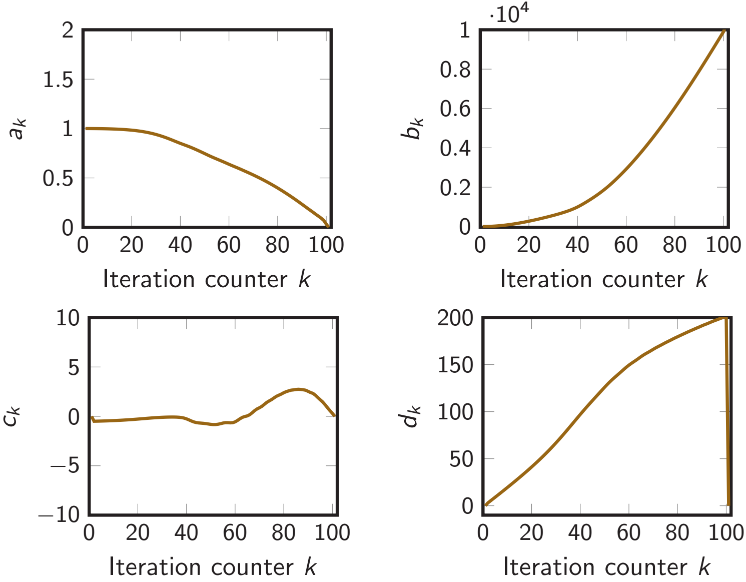
$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\left\|x_0-x_\star\right\|^2}{b_N}.$$

   Numerically (live if time allows)

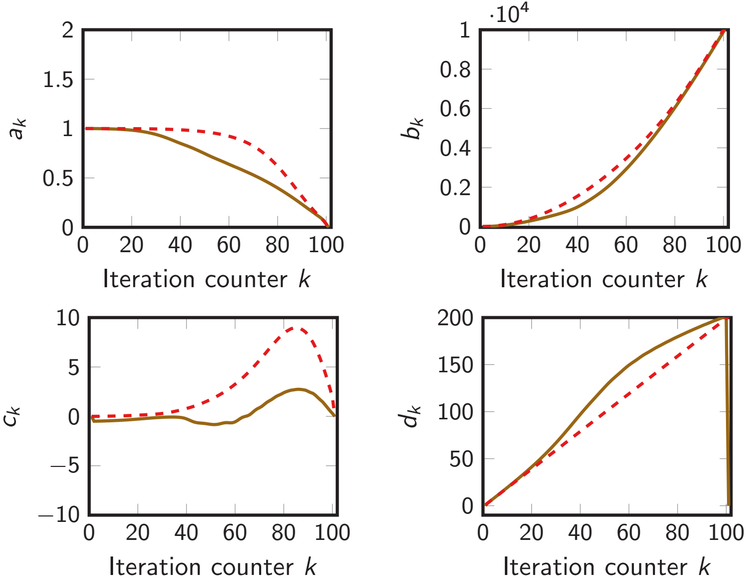
$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \dots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \dots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.
   Tentative simplification #1: $d_k = (2k+1)L$

$$V_k = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top \left[\begin{pmatrix} a_k & c_k \\ c_k & b_k \end{pmatrix} \otimes I_d \right] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + d_k \left(f(x_k) - f(x_\star) \right)$$



$$V_k = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top \left[\begin{pmatrix} a_k & c_k \\ c_k & b_k \end{pmatrix} \otimes I_d \right] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + (2k+1)L\left(f(x_k) - f(x_\star)\right)$$



How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\| f'(x_N) \right\|^2 \leqslant \frac{L^2 \, \|x_0 - x_\star\|^2}{b_N}.$$

   Numerically (live if time allows)

$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \ldots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \ldots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.

   Tentative simplification #1: $d_k = (2k+1)L$

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\left\|x_0 - x_\star\right\|^2}{b_N}.$$

   Numerically (live if time allows)

$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \ldots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \ldots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.

   Tentative simplification #1: $d_k = (2k+1)L$ [it works!]

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

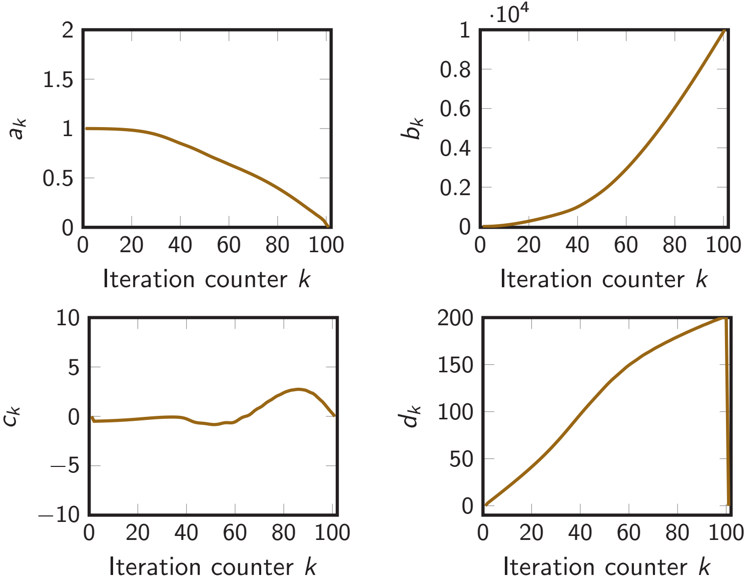
$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2 \|x_0 - x_\star\|^2}{b_N}.$$

   Numerically (live if time allows)

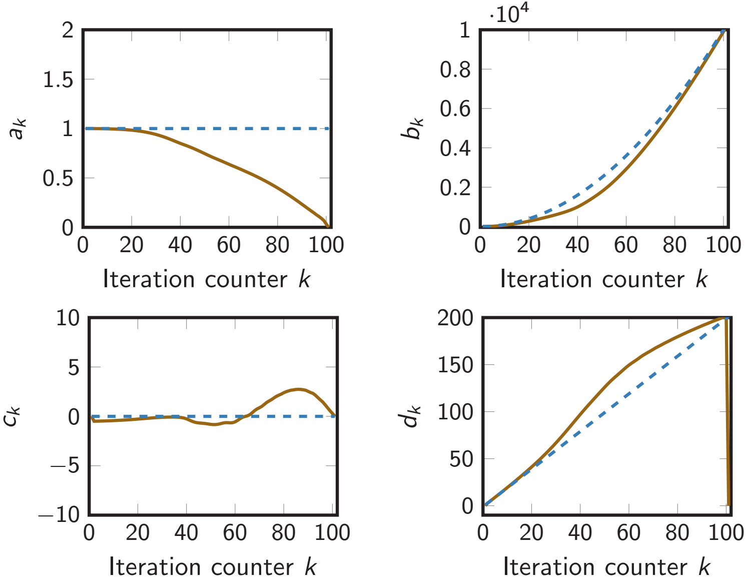
$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \ldots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \ldots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.
   Tentative simplification #1: $d_k = (2k+1)L$ [it works!]
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$

$$V_k = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top \left[\begin{pmatrix} a_k & c_k \\ c_k & b_k \end{pmatrix} \otimes I_d \right] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + (2k+1)L\left(f(x_k) - f(x_\star)\right)$$



$$V_k = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top \left[\begin{pmatrix} L^2 & 0 \\ 0 & b_k \end{pmatrix} \otimes I_d \right] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + (2k+1)L\left(f(x_k) - f(x_\star)\right)$$



How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\,\|x_0-x_\star\|^2}{b_N}.$$

   Numerically (live if time allows)

$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \dots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \dots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.
   Tentative simplification #1: $d_k = (2k+1)L$ [it works!]
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2 \left\|x_0 - x_\star\right\|^2}{b_N}.$$

   Numerically (live if time allows)

$$\begin{array}{rccccccc} N = & 1 & 2 & 3 & 4 & \ldots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \ldots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.
   Tentative simplification #1: $d_k = (2k+1)L$ [it works!]
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$ [it works!]

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

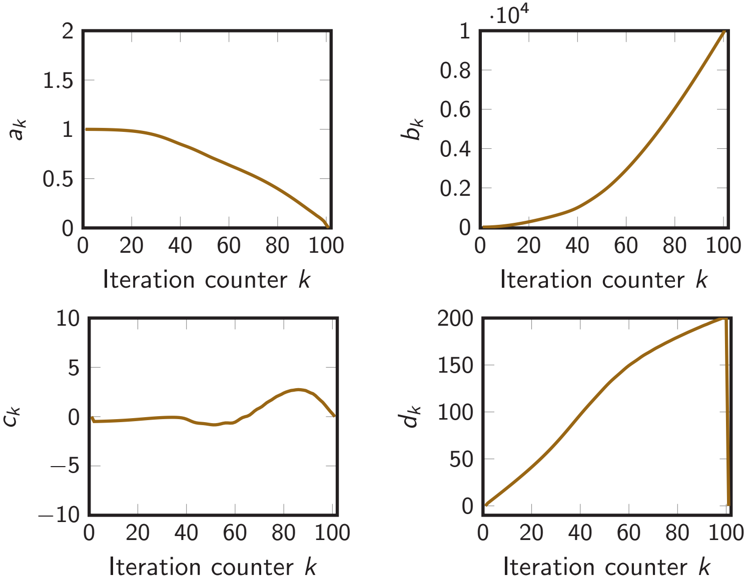
$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\,\|x_0-x_\star\|^2}{b_N}.$$

Numerically (live if time allows)

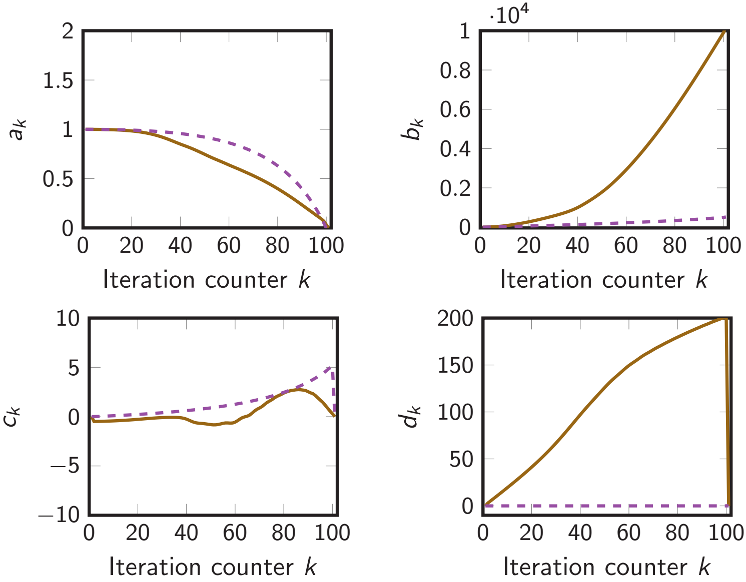
$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \dots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \dots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.
   Tentative simplification #1: $d_k = (2k+1)L$ [it works!]
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$ [it works!]
   Tentative simplification #3: $d_k = 0$

$$V_k = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top \left[\begin{pmatrix} a_k & c_k \\ c_k & b_k \end{pmatrix} \otimes I_d \right] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + d_k \left(f(x_k) - f(x_\star) \right)$$



$$V_k = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top \left[\begin{pmatrix} a_k & c_k \\ c_k & b_k \end{pmatrix} \otimes I_d \right] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + 0\,(f(x_k) - f(x_\star))$$



How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\| f'(x_N) \right\|^2 \leqslant \frac{L^2 \left\| x_0 - x_\star \right\|^2}{b_N}.$$

   Numerically (live if time allows)

$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \ldots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \ldots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.
   Tentative simplification #1: $d_k = (2k+1)L$ [it works!]
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$ [it works!]
   Tentative simplification #3: $d_k = 0$

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2\,\|x_0 - x_\star\|^2}{b_N}.$$

Numerically (live if time allows)

$$\begin{array}{rccccccc} N = & 1 & 2 & 3 & 4 & \dots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \dots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.
   Tentative simplification #1: $d_k = (2k+1)L$ [it works!]
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$ [it works!]
   Tentative simplification #3: $d_k = 0$ [it fails!]

How does it work for the gradient method?

1. Solve the SDP for some values of N; recall final guarantee of the form:

$$\left\|f'(x_N)\right\|^2 \leqslant \frac{L^2 \|x_0 - x_\star\|^2}{b_N}.$$

   Numerically (live if time allows)

$$\begin{array}{rccccccc} N = & 1 & 2 & 3 & 4 & \ldots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \ldots & 10201 \end{array}$$

2. Observe the a_k, b_k, c_k, d_k's for some values of N.
3. Try to simplify the ϕ_k^f's without loosing too much.
   Tentative simplification #1: $d_k = (2k+1)L$ [it works!]
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$ [it works!]
   Tentative simplification #3: $d_k = 0$ [it fails!]
4. Prove target result by analytically playing with $\mathcal{V}_k$:

$$\phi_k^f(x_k) = (2k+1)L(f(x_k) - f_\star) + k(k+2)\left\|f'(x_k)\right\|^2 + L^2\|x_k - x_\star\|^2,$$

   hence $f(x_N) - f_\star = O(N^{-1})$ and $\|f'(x_N)\|^2 = O(N^{-2})$ using $b_N = N(N+2)$.

Remaining questions

From previous content, we should still answer

- ◇ how to obtain a suitable representation of $\mathcal{V}_k$?
- ◇ How to obtain an analytical potential, rigorously?
- ◇ Does it apply beyond gradient descent?

Toy example: gradient descent

Reformulation as a LMI

Other examples

Concluding remarks

Toy example: gradient descent

Reformulation as a LMI

Other examples

Concluding remarks

Verifying potentials

Verifying potentials

Recall our candidate quadratic ϕ_k^f with *all the available information* at iteration k

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k \left(f(x_k) - f_\star\right).$$

Verifying potentials

Recall our candidate quadratic ϕ_k^f with *all the available information* at iteration k

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \|f'(x_k)\|^2 + 2c_k \langle f'(x_k), x_k - x_\star \rangle + d_k (f(x_k) - f_\star).$$

Given ϕ_{k+1}^f, ϕ_k^f (i.e., fixed $\{a_k, b_k, c_k, d_k\}$), *how to verify*

$$\phi_{k+1}^f \leq \phi_k^f$$

for all L-smooth convex f and x_k?

Verifying potentials

Recall our candidate quadratic ϕ_k^f with *all the available information* at iteration k

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k \left(f(x_k) - f_\star\right).$$

Given ϕ_{k+1}^f, ϕ_k^f (i.e., fixed $\{a_k, b_k, c_k, d_k\}$), *how to verify*

$$\phi_{k+1}^f \leq \phi_k^f$$

for all L-smooth convex f and x_k?

Base idea: reformulate question as verification of

$$0 \geq \qquad \phi_{k+1}^f - \phi_k^f \qquad ,$$

Verifying potentials

Recall our candidate quadratic ϕ_k^f with *all the available information* at iteration k

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \|f'(x_k)\|^2 + 2c_k \langle f'(x_k), x_k - x_\star \rangle + d_k (f(x_k) - f_\star).$$

Given ϕ_{k+1}^f, ϕ_k^f (i.e., fixed $\{a_k, b_k, c_k, d_k\}$), *how to verify*

$$\phi_{k+1}^f \leq \phi_k^f$$

for all L-smooth convex f and x_k?

Base idea: reformulate question as verification of

$$0 \geq \max_{x_k, x_{k+1}, f} \phi_{k+1}^f - \phi_k^f \qquad ,$$

Verifying potentials

Recall our candidate quadratic ϕ_k^f with *all the available information* at iteration k

$$\phi_k^f = a_k \left\|x_k - x_\star\right\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \left\langle f'(x_k), x_k - x_\star \right\rangle + d_k \left(f(x_k) - f_\star\right).$$

Given ϕ_{k+1}^f, ϕ_k^f (i.e., fixed $\{a_k, b_k, c_k, d_k\}$), *how to verify*

$$\phi_{k+1}^f \leq \phi_k^f$$

for all L-smooth convex f and x_k?

Base idea: reformulate question as verification of

$$0 \geq \max_{x_k, x_{k+1}, f} \phi_{k+1}^f - \phi_k^f \text{ s.t. } f \text{ convex and } L\text{-smooth}, x_{k+1} = x_k - \gamma_k f'(x_k),$$

Verifying potentials

Recall our candidate quadratic ϕ_k^f with *all the available information* at iteration k

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \langle f'(x_k), x_k - x_\star\rangle + d_k (f(x_k) - f_\star).$$

Given ϕ_{k+1}^f, ϕ_k^f (i.e., fixed $\{a_k, b_k, c_k, d_k\}$), *how to verify*

$$\phi_{k+1}^f \leq \phi_k^f$$

for all L-smooth convex f and x_k?

Base idea: reformulate question as verification of

$$0 \geq \max_{x_k, x_{k+1}, f} \phi_{k+1}^f - \phi_k^f \text{ s.t. } f \text{ convex and } L\text{-smooth}, x_{k+1} = x_k - \gamma_k f'(x_k),$$

i.e.: replace "for all" by maximization (later formulated as a *semidefinite program*).

Verifying potentials

For exposition purposes, let us treat the simpler case $\phi_k^f = a_k \left\| x_k - x_\star \right\|^2$.

Verifying potentials

For exposition purposes, let us treat the simpler case $\phi_k^f = a_k \|x_k - x_\star\|^2$.

Verifying $\phi_{k+1}^f \le \phi_k^f$ (for all f and x_k) is equivalent to verify

$$\begin{aligned} 0 \ge \max_{x_k, x_{k+1}, f} \; & a_{k+1} \|x_{k+1} - x_\star\|^2 - a_k \|x_k - x_\star\|^2 \\ \text{subject to } & f \text{ is convex and } L\text{-smooth}, \\ & x_{k+1} = x_k - \gamma_k f'(x_k). \end{aligned}$$

Verifying potentials

For exposition purposes, let us treat the simpler case $\phi_k^f = a_k \|x_k - x_\star\|^2$.

Verifying $\phi_{k+1}^f \leq \phi_k^f$ (for all f and x_k) is equivalent to verify

$$\begin{aligned} 0 \geq \max_{x_k, x_{k+1}, f} \; & a_{k+1} \|x_{k+1} - x_\star\|^2 - a_k \|x_k - x_\star\|^2 \\ \text{subject to } & f \text{ is convex and } L\text{-smooth}, \\ & x_{k+1} = x_k - \gamma_k f'(x_k). \end{aligned}$$

This is an *infinite dimensional* problem (variables: f, x_k and x_{k+1}).

First simplification: sampling

As it is, the previous problem does not seem very practical...

First simplification: sampling

As it is, the previous problem does not seem very practical...

- How to treat the infinite dimensional variable f?

First simplification: sampling

As it is, the previous problem does not seem very practical...

- How to treat the infinite dimensional variable f?
- How to cope with the constraint f convex and L-smooth?

First simplification: sampling

As it is, the previous problem does not seem very practical...

- How to treat the infinite dimensional variable f?
- How to cope with the constraint f convex and L-smooth?

Idea:

First simplification: sampling

As it is, the previous problem does not seem very practical...

- How to treat the infinite dimensional variable f?
- How to cope with the constraint f convex and L-smooth?

Idea:

- replace f by its discrete version:

$$f_i = f(x_i),\ g_i = f'(x_i) \quad \forall i \in \{k, \star\}.$$

First simplification: sampling

As it is, the previous problem does not seem very practical...

- How to treat the infinite dimensional variable f?
- How to cope with the constraint f convex and L-smooth?

Idea:

- replace f by its discrete version:

$$f_i = f(x_i),\ g_i = f'(x_i) \quad \forall i \in \{k, \star\}\,.$$

- Require points (x_i, g_i, f_i) to be interpolable by a convex and L-smooth f.

First simplification: sampling

As it is, the previous problem does not seem very practical...

- How to treat the infinite dimensional variable f?
- How to cope with the constraint f convex and L-smooth?

Idea:

- replace f by its discrete version:

$$f_i = f(x_i),\ g_i = f'(x_i) \quad \forall i \in \{k, \star\}\,.$$

- Require points (x_i, g_i, f_i) to be interpolable by a convex and L-smooth f.
  The new constraint is:

$$\exists f \text{ (convex and } L\text{-smooth)}:\ f_i = f(x_i),\ g_i = f'(x_i), \qquad \forall i \in \{k, \star\}\,.$$

Verifying potentials

For exposition purposes, let us treat the simpler case $\phi_k^f = a_k \|x_k - x_\star\|^2$.

Verifying $\phi_{k+1}^f \leq \phi_k^f$ (for all f and x_k) is equivalent to

$$0 \geq \max_{x_k, x_{k+1}, f} a_{k+1} \|x_{k+1} - x_\star\|^2 - a_k \|x_k - x_\star\|^2$$

$$\text{subject to } f \text{ is convex and } L\text{-smooth,}$$

$$x_{k+1} = x_k - \gamma_k f'(x_k).$$

This is an *infinite dimensional* problem (variables: f, x_k and x_{k+1}).

Verifying potentials

For exposition purposes, let us treat the simpler case $\phi_k^f = a_k \|x_k - x_\star\|^2$.

Verifying $\phi_{k+1}^f \leq \phi_k^f$ (for all f and x_k) is equivalent to

$$\begin{aligned} 0 \geq \max_{x_k, x_{k+1}, f} \; & a_{k+1} \|x_{k+1} - x_\star\|^2 - a_k \|x_k - x_\star\|^2 \\ \text{subject to } & f \text{ is convex and } L\text{-smooth}, \\ & x_{k+1} = x_k - \gamma_k f'(x_k). \end{aligned}$$

This is an *infinite dimensional* problem (variables: f, x_k and x_{k+1}).

Sampling: f and f' are evaluated only at x_k and $x_\star$, hence equivalent

$$\begin{aligned} 0 \geq \max_{x_\star, x_k, g_k, f_\star, f_k} \; & a_{k+1} \|x_k - \gamma_k g_k - x_\star\|^2 - a_k \|x_k - x_\star\|^2 \\ \text{subject to } & \exists f \text{ convex and } L\text{-smooth} : \begin{cases} g_k = f'(x_k), & f_k = f(x_k) \\ 0 = f'(x_\star), & f_\star = f(x_\star). \end{cases} \end{aligned}$$

Verifying potentials

For exposition purposes, let us treat the simpler case $\phi_k^f = a_k \|x_k - x_\star\|^2$.

Verifying $\phi_{k+1}^f \leq \phi_k^f$ (for all f and x_k) is equivalent to

$$0 \geq \max_{x_k, x_{k+1}, f} a_{k+1} \|x_{k+1} - x_\star\|^2 - a_k \|x_k - x_\star\|^2$$

$$\text{subject to } f \text{ is convex and } L\text{-smooth,}$$

$$x_{k+1} = x_k - \gamma_k f'(x_k).$$

This is an *infinite dimensional* problem (variables: f, x_k and x_{k+1}).

Sampling: f and f' are evaluated only at x_k and $x_\star$, hence equivalent

$$0 \geq \max_{x_\star, x_k, g_k, f_\star, f_k} a_{k+1} \|x_k - \gamma_k g_k - x_\star\|^2 - a_k \|x_k - x_\star\|^2$$

$$\text{subject to } \exists f \text{ convex and } L\text{-smooth} : \begin{cases} g_k = f'(x_k), & f_k = f(x_k) \\ 0 = f'(x_\star), & f_\star = f(x_\star). \end{cases}$$

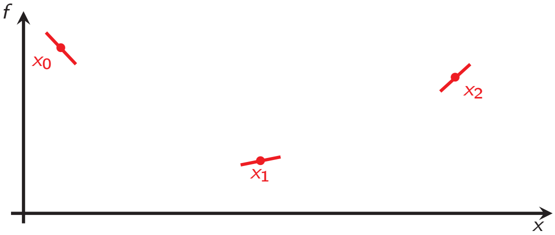
new variables: x_k, $x_\star$, g_k, $f_\star$, f_k. How to handle the existence constraint?

Smooth convex interpolation

Consider an index set S, and its associated values $\{(x_i, g_i, f_i)\}_{i \in S}$ with coordinates x_i, gradients g_i and function values f_i.

Smooth convex interpolation

Consider an index set S, and its associated values $\{(x_i, g_i, f_i)\}_{i\in S}$ with coordinates x_i, gradients g_i and function values f_i.

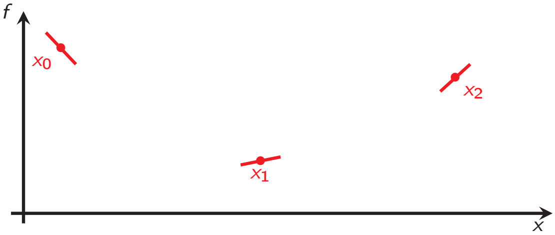


? Possible to find f convex and L-smooth such that

$$f(x_i) = f_i, \quad \text{and} \quad g_i = f'(x_i), \qquad \forall i \in S.$$

Smooth convex interpolation

Consider an index set S, and its associated values $\{(x_i, g_i, f_i)\}_{i\in S}$ with coordinates x_i, gradients g_i and function values f_i.



? Possible to find f convex and L-smooth such that

$$f(x_i) = f_i, \quad \text{and} \quad g_i = f'(x_i), \qquad \forall i \in S.$$

- Necessary and sufficient condition: $\forall i, j \in S$

$$f_i \geqslant f_j + \langle g_j, x_i - x_j \rangle + \frac{1}{2L}\|g_i - g_j\|^2.$$

Quadratic reformulation

From sampling, we had "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k) iff

$$0 \geq \max_{x_\star, x_k, g_k, f_\star, f_k} a_{k+1} \left\|x_k - \gamma_k g_k - x_\star\right\|^2 - a_k \left\|x_k - x_\star\right\|^2$$

$$\text{subject to } \exists f \text{ convex and } L\text{-smooth} : \begin{cases} g_k = f'(x_k), & f_k = f(x_k) \\ 0 = f'(x_\star), & f_\star = f(x_\star). \end{cases}$$

Quadratic reformulation

From sampling, we had "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k) iff

$$0 \geq \max_{x_\star, x_k, g_k, f_\star, f_k} a_{k+1} \left\|x_k - \gamma_k g_k - x_\star\right\|^2 - a_k \left\|x_k - x_\star\right\|^2$$

$$\text{subject to } \exists f \text{ convex and } L\text{-smooth} : \begin{cases} g_k = f'(x_k), & f_k = f(x_k) \\ 0 = f'(x_\star), & f_\star = f(x_\star). \end{cases}$$

and we can replace the *existence constraints* by

$$\begin{aligned} f_k &\geq f_\star + \tfrac{1}{2L}\|g_k\|^2, \\ f_\star &\geq f_k + \langle g_k, x_\star - x_k\rangle + \tfrac{1}{2L}\|g_k\|^2, \end{aligned}$$

reaching a (nonconvex) quadratic problem.

Semidefinite reformulation

Quadratic reformulation: "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k) iff

$$\begin{aligned} 0 \geq \max_{x_\star, x_k, g_k, f_\star, f_k} \; & a_{k+1} \left\| x_k - \gamma_k g_k - x_\star \right\|^2 - a_k \left\| x_k - x_\star \right\|^2 \\ \text{subject to } & f_k \geq f_\star + \tfrac{1}{2L} \|g_k\|^2, \\ & f_\star \geq f_k + \langle g_k, x_\star - x_k \rangle + \tfrac{1}{2L} \|g_k\|^2, \end{aligned}$$

Semidefinite reformulation

Quadratic reformulation: "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k) iff

$$
\begin{aligned}
0 \geq \max_{x_\star, x_k, g_k, f_\star, f_k} \; & a_{k+1} \|x_k - \gamma_k g_k - x_\star\|^2 - a_k \|x_k - x_\star\|^2 \\
\text{subject to } \; & f_k \geq f_\star + \tfrac{1}{2L}\|g_k\|^2, \\
& f_\star \geq f_k + \langle g_k, x_\star - x_k \rangle + \tfrac{1}{2L}\|g_k\|^2,
\end{aligned}
$$

which is linear in terms of

$$
G = \begin{bmatrix} \|x_k - x_\star\|^2 & \langle x_k - x_\star, g_k \rangle \\ \langle x_k - x_\star, g_k \rangle & \|g_k\|^2 \end{bmatrix}, \quad F = \begin{bmatrix} f_k & f_\star \end{bmatrix},
$$

where $G \succcurlyeq 0$ by construction.

Semidefinite reformulation

Quadratic reformulation: "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k) iff

$$
\begin{aligned}
0 \geq \max_{x_\star, x_k, g_k, f_\star, f_k} \; & a_{k+1} \|x_k - \gamma_k g_k - x_\star\|^2 - a_k \|x_k - x_\star\|^2 \\
\text{subject to } \; & f_k \geq f_\star + \tfrac{1}{2L} \|g_k\|^2, \\
& f_\star \geq f_k + \langle g_k, x_\star - x_k \rangle + \tfrac{1}{2L} \|g_k\|^2,
\end{aligned}
$$

which is linear in terms of

$$
G = \begin{bmatrix} \|x_k - x_\star\|^2 & \langle x_k - x_\star, g_k \rangle \\ \langle x_k - x_\star, g_k \rangle & \|g_k\|^2 \end{bmatrix}, \quad F = \begin{bmatrix} f_k & f_\star \end{bmatrix},
$$

where $G \succcurlyeq 0$ by construction. Hence "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k) iff

$$
\begin{aligned}
0 \geq \max_{G \succcurlyeq 0,\, F} \; & a_{k+1} (G_{1,1} + \gamma_k^2 G_{2,2} - 2\gamma_k G_{1,2}) - a_k \, G_{1,1} \\
\text{subject to } \; & F_1 \geq F_2 + \tfrac{1}{2L} G_{2,2}, \\
& F_2 \geq F_1 + G_{1,2} + \tfrac{1}{2L} G_{2,2},
\end{aligned}
$$

Semidefinite reformulation

Quadratic reformulation: "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k) iff

$$
\begin{aligned}
0 \geq \max_{x_\star, x_k, g_k, f_\star, f_k} \; & a_{k+1} \|x_k - \gamma_k g_k - x_\star\|^2 - a_k \|x_k - x_\star\|^2 \\
\text{subject to } \; & f_k \geq f_\star + \tfrac{1}{2L}\|g_k\|^2, \\
& f_\star \geq f_k + \langle g_k, x_\star - x_k\rangle + \tfrac{1}{2L}\|g_k\|^2,
\end{aligned}
$$

which is linear in terms of

$$
G = \begin{bmatrix} \|x_k - x_\star\|^2 & \langle x_k - x_\star, g_k\rangle \\ \langle x_k - x_\star, g_k\rangle & \|g_k\|^2 \end{bmatrix}, \quad F = \begin{bmatrix} f_k & f_\star \end{bmatrix},
$$

where $G \succcurlyeq 0$ by construction. Hence "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k) iff

$$
\begin{aligned}
0 \geq \max_{G \succcurlyeq 0,\, F} \; & a_{k+1} \left(G_{1,1} + \gamma_k^2 G_{2,2} - 2\gamma_k G_{1,2}\right) - a_k\, G_{1,1} \\
\text{subject to } \; & F_1 \geq F_2 + \tfrac{1}{2L} G_{2,2}, \\
& F_2 \geq F_1 + G_{1,2} + \tfrac{1}{2L} G_{2,2},
\end{aligned}
$$

which is a regular *semidefinite program* (SDP).

Verifying a potential

Final step: inequality verified iff *dual SDP is feasible*, that is

$$
\begin{aligned}
0 \geq \max_{G \succcurlyeq 0,\, F} &\; a_{k+1}\,(G_{1,1} + \gamma_k^2 G_{2,2} - 2G_{1,2}) - a_k\, G_{1,1} \\
\text{subject to } &\; F_1 \geq F_2 + \tfrac{1}{2L} G_{2,2} && : \lambda_1, \\
&\; F_2 \geq F_1 + G_{1,2} + \tfrac{1}{2L} G_{2,2} && : \lambda_2.
\end{aligned}
$$

Verifying a potential

Final step: inequality verified iff *dual SDP is feasible*, that is

$$
\begin{aligned}
0 \geq \max_{G \succcurlyeq 0,\, F} & \; a_{k+1}\,(G_{1,1} + \gamma_k^2 G_{2,2} - 2G_{1,2}) - a_k\, G_{1,1} & \\
\text{subject to } & F_1 \geq F_2 + \tfrac{1}{2L} G_{2,2} & : \lambda_1, \\
& F_2 \geq F_1 + G_{1,2} + \tfrac{1}{2L} G_{2,2} & : \lambda_2.
\end{aligned}
$$

The dual problem has the form (note that no duality gap occurs):

$$
\begin{aligned}
0 \geq \min_{\lambda_1, \lambda_2 \geq 0} & \; 0 \\
\text{subject to } & \lambda_1 = \lambda_2, \\
& \begin{pmatrix} a_k - a_{k+1} & \gamma_k a_{k+1} - \frac{\lambda_2}{2} \\ \gamma_k a_{k+1} - \frac{\lambda_2}{2} & \frac{1}{2L}(\lambda_1 + \lambda_2) - a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0,
\end{aligned}
$$

Verifying a potential

Final step: inequality verified iff *dual SDP is feasible*, that is

$$\begin{aligned} 0 \geq \max_{G \succcurlyeq 0,\, F} \; & a_{k+1}\left(G_{1,1} + \gamma_k^2 G_{2,2} - 2G_{1,2}\right) - a_k\, G_{1,1} \\ \text{subject to } & F_1 \geq F_2 + \tfrac{1}{2L} G_{2,2} && : \lambda_1, \\ & F_2 \geq F_1 + G_{1,2} + \tfrac{1}{2L} G_{2,2} && : \lambda_2. \end{aligned}$$

The dual problem has the form (note that no duality gap occurs):

$$\begin{aligned} 0 \geq \min_{\lambda_1, \lambda_2 \geq 0} \; & 0 \\ \text{subject to } & \lambda_1 = \lambda_2, \\ & \begin{pmatrix} a_k - a_{k+1} & \gamma_k a_{k+1} - \frac{\lambda_2}{2} \\ \gamma_k a_{k+1} - \frac{\lambda_2}{2} & \frac{1}{2L}(\lambda_1 + \lambda_2) - a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0, \end{aligned}$$

hence *feasibility problem* equivalent to verification "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k).

Verifying a potential: final formulation

$$a_{k+1}\|x_{k+1}-x_\star\|^2 \leq a_k\|x_k-x_\star\|^2 \text{ with } x_{k+1}=x_k-\gamma_k f'(x_k)\text{, for all } L\text{-smooth convex } f \text{ and } x_k$$

$$\Leftrightarrow$$

$$\exists\lambda\geq 0: \quad \begin{pmatrix} a_k-a_{k+1} & \gamma_k a_{k+1}-\frac{\lambda}{2} \\ \gamma_k a_{k+1}-\frac{\lambda}{2} & \frac{\lambda}{L}-a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0.$$

Verifying a potential: final formulation

$a_{k+1}\|x_{k+1}-x_\star\|^2 \leq a_k\|x_k-x_\star\|^2$ with $x_{k+1}=x_k-\gamma_k f'(x_k)$, for all L-smooth convex f and x_k

$$\Leftrightarrow$$

$$\exists\lambda\geq 0: \quad \begin{pmatrix} a_k-a_{k+1} & \gamma_k a_{k+1}-\frac{\lambda}{2} \\ \gamma_k a_{k+1}-\frac{\lambda}{2} & \frac{\lambda}{L}-a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0.$$

How to verify more complicated potential, such as

$$\phi_k^f = a_k\left\|x_k-x_\star\right\|^2 + b_k\left\|f'(x_k)\right\|^2 + 2c_k\left\langle f'(x_k), x_k-x_\star\right\rangle + d_k\left(f(x_k)-f_\star\right)?$$

Verifying a potential: final formulation

$$a_{k+1}\|x_{k+1}-x_\star\|^2 \leq a_k\|x_k-x_\star\|^2 \text{ with } x_{k+1}=x_k-\gamma_k f'(x_k), \text{ for all } L\text{-smooth convex } f \text{ and } x_k$$

$$\Leftrightarrow$$

$$\exists\lambda\geq 0: \quad \begin{pmatrix} a_k-a_{k+1} & \gamma_k a_{k+1}-\frac{\lambda}{2} \\ \gamma_k a_{k+1}-\frac{\lambda}{2} & \frac{\lambda}{L}-a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0.$$

How to verify more complicated potential, such as

$$\phi_k^f = a_k\left\|x_k-x_\star\right\|^2 + b_k\left\|f'(x_k)\right\|^2 + 2c_k\left\langle f'(x_k), x_k-x_\star\right\rangle + d_k\left(f(x_k)-f_\star\right)?$$

Exact *same tricks*, with some adaptations

Verifying a potential: final formulation

$a_{k+1}\|x_{k+1}-x_\star\|^2 \leq a_k\|x_k - x_\star\|^2$ with $x_{k+1} = x_k - \gamma_k f'(x_k)$, for all L-smooth convex f and x_k

$$\Leftrightarrow$$

$$\exists \lambda \geq 0 : \quad \begin{pmatrix} a_k - a_{k+1} & \gamma_k a_{k+1} - \frac{\lambda}{2} \\ \gamma_k a_{k+1} - \frac{\lambda}{2} & \frac{\lambda}{L} - a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0.$$

How to verify more complicated potential, such as

$$\phi_k^f = a_k \left\|x_k - x_\star\right\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \left\langle f'(x_k), x_k - x_\star\right\rangle + d_k \left(f(x_k) - f_\star\right)?$$

Exact *same tricks*, with some adaptations

- additional sample x_{k+1} (for using $g_{k+1} = f'(x_{k+1})$, $f_{k+1} = f(x_{k+1})$),

Verifying a potential: final formulation

$$a_{k+1}\|x_{k+1}-x_\star\|^2 \leq a_k\|x_k-x_\star\|^2 \text{ with } x_{k+1}=x_k-\gamma_k f'(x_k), \text{ for all } L\text{-smooth convex } f \text{ and } x_k$$

$$\Leftrightarrow$$

$$\exists\lambda\geq 0: \quad \begin{pmatrix} a_k-a_{k+1} & \gamma_k a_{k+1}-\frac{\lambda}{2} \\ \gamma_k a_{k+1}-\frac{\lambda}{2} & \frac{\lambda}{L}-a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0.$$

How to verify more complicated potential, such as

$$\phi_k^f = a_k\left\|x_k-x_\star\right\|^2 + b_k\left\|f'(x_k)\right\|^2 + 2c_k\left\langle f'(x_k), x_k-x_\star\right\rangle + d_k\left(f(x_k)-f_\star\right)?$$

Exact *same tricks*, with some adaptations

- additional sample x_{k+1} (for using $g_{k+1}=f'(x_{k+1})$, $f_{k+1}=f(x_{k+1})$),
- hence 6 inequalities (instead of 2),

Verifying a potential: final formulation

$$a_{k+1}\|x_{k+1}-x_\star\|^2 \leq a_k\|x_k-x_\star\|^2 \text{ with } x_{k+1}=x_k-\gamma_k f'(x_k), \text{ for all}$$
$$L\text{-smooth convex } f \text{ and } x_k$$
$$\Leftrightarrow$$
$$\exists\lambda\geq 0: \quad \begin{pmatrix} a_k-a_{k+1} & \gamma_k a_{k+1}-\frac{\lambda}{2} \\ \gamma_k a_{k+1}-\frac{\lambda}{2} & \frac{\lambda}{L}-a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0.$$

How to verify more complicated potential, such as

$$\phi_k^f = a_k\left\|x_k-x_\star\right\|^2 + b_k\left\|f'(x_k)\right\|^2 + 2c_k\left\langle f'(x_k), x_k-x_\star\right\rangle + d_k\left(f(x_k)-f_\star\right)?$$

Exact *same tricks*, with some adaptations

- additional sample x_{k+1} (for using $g_{k+1}=f'(x_{k+1})$, $f_{k+1}=f(x_{k+1})$),
- hence 6 inequalities (instead of 2),
- and 3x3 SDP (also on dual side).

Verifying a potential: final formulation

$a_{k+1}\|x_{k+1}-x_\star\|^2 \leq a_k\|x_k-x_\star\|^2$ with $x_{k+1}=x_k-\gamma_k f'(x_k)$, for all L-smooth convex f and x_k

$$\Leftrightarrow$$

$$\exists \lambda \geq 0: \quad \begin{pmatrix} a_k - a_{k+1} & \gamma_k a_{k+1} - \frac{\lambda}{2} \\ \gamma_k a_{k+1} - \frac{\lambda}{2} & \frac{\lambda}{L} - a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0.$$

How to verify more complicated potential, such as

$$\phi_k^f = a_k \left\|x_k - x_\star\right\|^2 + b_k \left\|f'(x_k)\right\|^2 + 2c_k \left\langle f'(x_k), x_k - x_\star \right\rangle + d_k \left(f(x_k) - f_\star\right)?$$

Exact *same tricks*, with some adaptations

- additional sample x_{k+1} (for using $g_{k+1} = f'(x_{k+1})$, $f_{k+1} = f(x_{k+1})$),
- hence 6 inequalities (instead of 2),
- and 3x3 SDP (also on dual side).

How to find a proof for "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k)?

Verifying a potential: final formulation

$a_{k+1}\|x_{k+1}-x_\star\|^2 \leq a_k\|x_k-x_\star\|^2$ with $x_{k+1}=x_k-\gamma_k f'(x_k)$, for all L-smooth convex f and x_k

$$\Leftrightarrow$$

$$\exists\lambda\geq 0: \quad \begin{pmatrix} a_k-a_{k+1} & \gamma_k a_{k+1}-\frac{\lambda}{2} \\ \gamma_k a_{k+1}-\frac{\lambda}{2} & \frac{\lambda}{L}-a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0.$$

How to verify more complicated potential, such as

$$\phi_k^f = a_k\left\|x_k-x_\star\right\|^2 + b_k\left\|f'(x_k)\right\|^2 + 2c_k\left\langle f'(x_k), x_k-x_\star\right\rangle + d_k\left(f(x_k)-f_\star\right)?$$

Exact *same tricks*, with some adaptations

- additional sample x_{k+1} (for using $g_{k+1}=f'(x_{k+1})$, $f_{k+1}=f(x_{k+1})$),
- hence 6 inequalities (instead of 2),
- and 3x3 SDP (also on dual side).

How to find a proof for "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k)?

- Exhibit a dual feasible point,

Verifying a potential: final formulation

$$a_{k+1}\|x_{k+1}-x_\star\|^2 \leq a_k\|x_k-x_\star\|^2 \text{ with } x_{k+1}=x_k-\gamma_k f'(x_k), \text{ for all}$$
$$L\text{-smooth convex } f \text{ and } x_k$$
$$\Leftrightarrow$$
$$\exists \lambda \geq 0: \quad \begin{pmatrix} a_k - a_{k+1} & \gamma_k a_{k+1} - \frac{\lambda}{2} \\ \gamma_k a_{k+1} - \frac{\lambda}{2} & \frac{\lambda}{L} - a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0.$$

How to verify more complicated potential, such as

$$\phi_k^f = a_k\|x_k-x_\star\|^2 + b_k\left\|f'(x_k)\right\|^2 + 2c_k\langle f'(x_k), x_k-x_\star\rangle + d_k\left(f(x_k)-f_\star\right)?$$

Exact *same tricks*, with some adaptations

- additional sample x_{k+1} (for using $g_{k+1}=f'(x_{k+1})$, $f_{k+1}=f(x_{k+1})$),
- hence 6 inequalities (instead of 2),
- and 3x3 SDP (also on dual side).

How to find a proof for "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k)?

- Exhibit a dual feasible point,
- proof only consists in combining quadratic inequalities.

Verifying a potential: final formulation

$$a_{k+1}\|x_{k+1}-x_\star\|^2 \leq a_k\|x_k-x_\star\|^2 \text{ with } x_{k+1}=x_k-\gamma_k f'(x_k), \text{ for all}$$
$$L\text{-smooth convex } f \text{ and } x_k$$
$$\Leftrightarrow$$
$$\exists\lambda\geq 0: \quad \begin{pmatrix} a_k-a_{k+1} & \gamma_k a_{k+1}-\frac{\lambda}{2} \\ \gamma_k a_{k+1}-\frac{\lambda}{2} & \frac{\lambda}{L}-a_{k+1}\gamma_k^2 \end{pmatrix} \succcurlyeq 0.$$

How to verify more complicated potential, such as

$$\phi_k^f = a_k\,\|x_k-x_\star\|^2 + b_k\,\left\|f'(x_k)\right\|^2 + 2c_k\,\langle f'(x_k), x_k-x_\star\rangle + d_k\,(f(x_k)-f_\star)?$$

Exact *same tricks*, with some adaptations

- additional sample x_{k+1} (for using $g_{k+1}=f'(x_{k+1})$, $f_{k+1}=f(x_{k+1})$),
- hence 6 inequalities (instead of 2),
- and 3x3 SDP (also on dual side).

How to find a proof for "$\phi_{k+1}^f \leq \phi_k^f$" (for all f and x_k)?

- Exhibit a dual feasible point,
- proof only consists in combining quadratic inequalities.
- If inequality does not hold (for all f and x_k), primal solutions are counter-examples.

Toy example: gradient descent

Reformulation as a LMI

Other examples

Concluding remarks

Toy example: gradient descent

Reformulation as a LMI

Other examples

Concluding remarks

Accelerated/fast gradient method

Consider a fast gradient method for smooth convex minimization:

Accelerated/fast gradient method

Consider a fast gradient method for smooth convex minimization:

$$
\begin{aligned}
y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k, \\
x_{k+1} &= y_{k+1} - \alpha_k f'(y_{k+1}), \\
z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}),
\end{aligned}
$$

Accelerated/fast gradient method

Consider a fast gradient method for smooth convex minimization:

$$
\begin{aligned}
y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k, \\
x_{k+1} &= y_{k+1} - \alpha_k f'(y_{k+1}), \\
z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}),
\end{aligned}
$$

with carefully chosen values of τ_k and η_k.

Accelerated/fast gradient method

Consider a fast gradient method for smooth convex minimization:

$$
\begin{aligned}
y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k,\\
x_{k+1} &= y_{k+1} - \alpha_k f'(y_{k+1}),\\
z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}),
\end{aligned}
$$

with carefully chosen values of τ_k and η_k.

Same as before: pick a family of potentials, such as

$$
\phi_k^f = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top [Q_k \otimes I_d] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + a_k \|z_k - x_\star\|^2 + d_k \left(f(x_k) - f_\star\right),
$$

Accelerated/fast gradient method

Consider a fast gradient method for smooth convex minimization:

$$
\begin{aligned}
y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k,\\
x_{k+1} &= y_{k+1} - \alpha_k f'(y_{k+1}),\\
z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}),
\end{aligned}
$$

with carefully chosen values of τ_k and η_k.

Same as before: pick a family of potentials, such as

$$
\phi_k^f = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top [Q_k \otimes I_d] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + a_k \|z_k - x_\star\|^2 + d_k\,(f(x_k) - f_\star),
$$

and solve the corresponding SDP numerically:

$$
\max_{\phi_1^f,\ldots,\phi_{N-1}^f,d_N} d_N \text{ such that } (\phi_0^f,\phi_1^f) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f,\phi_N^f) \in \mathcal{V}_{N-1}.
$$

Accelerated/fast gradient method

Consider a fast gradient method for smooth convex minimization:

$$
\begin{aligned}
y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k, \\
x_{k+1} &= y_{k+1} - \alpha_k f'(y_{k+1}), \\
z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}),
\end{aligned}
$$

with carefully chosen values of τ_k and η_k.

Same as before: pick a family of potentials, such as

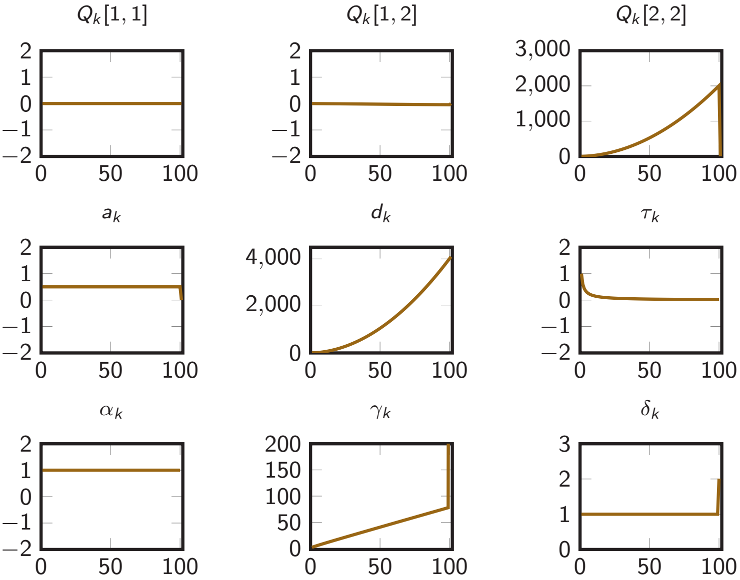
$$
\phi_k^f = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top [Q_k \otimes I_d] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + a_k \|z_k - x_\star\|^2 + d_k\,(f(x_k) - f_\star),
$$

and solve the corresponding SDP numerically:

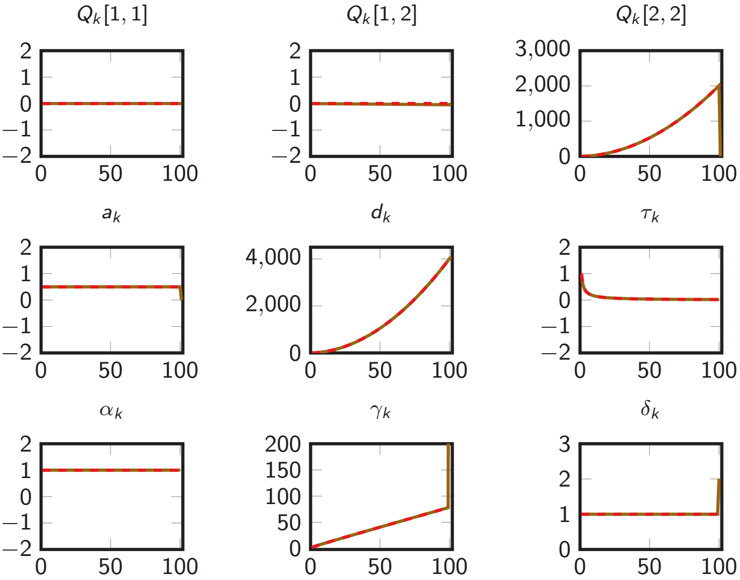
$$
\max_{\phi_1^f,\ldots,\phi_{N-1}^f,d_N} d_N \text{ such that } (\phi_0^f,\phi_1^f)\in\mathcal{V}_0,\ldots,(\phi_{N-1}^f,\phi_N^f)\in\mathcal{V}_{N-1}.
$$

Few additional technical ingredients allow tuning method's parameters simultaneously.

Example for $N = 100$, $L = 1$: numerics (brown),



Example for $N = 100$, $L = 1$: numerics (brown), and analytical solution (red).



Accelerated/fast gradient method

Consider a fast gradient method for smooth convex minimization:

$$
\begin{aligned}
y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k, \\
x_{k+1} &= y_{k+1} - \alpha_k f'(y_{k+1}), \\
z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}),
\end{aligned}
$$

with carefully chosen values of τ_k and η_k.

Accelerated/fast gradient method

Consider a fast gradient method for smooth convex minimization:

$$
\begin{aligned}
y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k,\\
x_{k+1} &= y_{k+1} - \alpha_k f'(y_{k+1}),\\
z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}),
\end{aligned}
$$

with carefully chosen values of τ_k and η_k.

Recovers standard potential (see e.g.,[Nesterov, 1983] or [Bansal & Gupta 2019]):

$$\phi_k^f = d_k(f(x_k) - f_\star) + \frac{L}{2}\|z_k - x_\star\|^2$$

with $d_k \sim k^2$ (more precisely: $d_{k+1} = 1 + d_k + \sqrt{1+d_k}$).

Accelerated/fast gradient method

Consider a fast gradient method for smooth convex minimization:

$$
\begin{aligned}
y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k,\\
x_{k+1} &= y_{k+1} - \alpha_k f'(y_{k+1}),\\
z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}),
\end{aligned}
$$

with carefully chosen values of τ_k and η_k.

Recovers standard potential (see e.g.,[Nesterov, 1983] or [Bansal & Gupta 2019]):

$$\phi_k^f = d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|z_k - x_\star\|^2$$

with $d_k \sim k^2$ (more precisely: $d_{k+1} = 1 + d_k + \sqrt{1+d_k}$).

From numerical inspirations, alternate ones also possible, such as

$$\phi_k^f = d_k'(f(x_k) - f_\star) + \tfrac{d_k'}{2L}\left\|f'(x_k)\right\|^2 + \tfrac{L}{2}\|z_k - x_\star\|^2$$

with $d_k' \sim k^2$ (more precisely: $d_{k+1}' = 1 + d_k' + \sqrt{1 + \tfrac{3}{2}d_k'}$, *red curves on prev. slide*).

Optimized gradient method

Optimized gradient methods (Kim & Fessler, 2016) can be factorized in a similar form

$$\begin{aligned} y_{k+1} &= (1-\tau_k) y_k + \tau_k z_k - \alpha_k f'(y_k), \\ z_{k+1} &= (1-\delta_k) y_{k+1} + \delta_k z_k - \gamma_k f'(y_k) - \gamma_k' f'(y_{k+1}), \end{aligned}$$

but current analysis is more involved (not based on potentials).

Optimized gradient method

Optimized gradient methods (Kim & Fessler, 2016) can be factorized in a similar form

$$\begin{aligned} y_{k+1} &= (1-\tau_k)y_k + \tau_k z_k - \alpha_k f'(y_k), \\ z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_k) - \gamma_k' f'(y_{k+1}), \end{aligned}$$

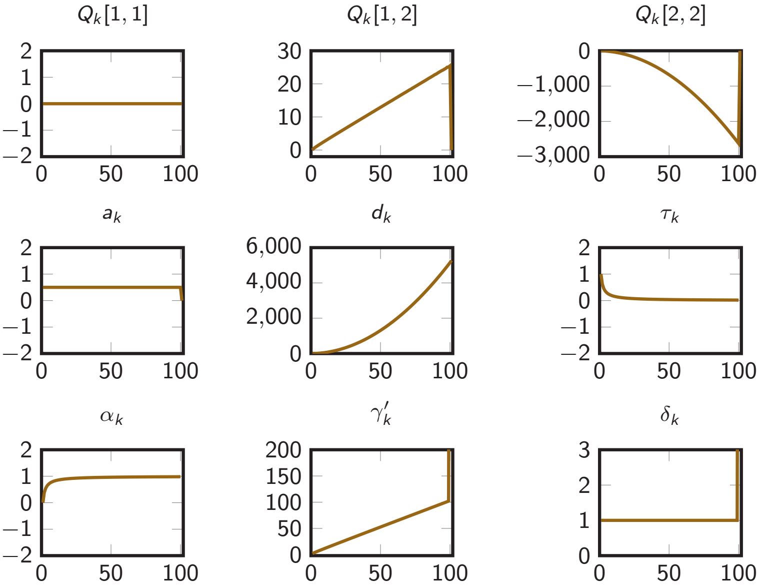
but current analysis is more involved (not based on potentials).

Starting with, for example,

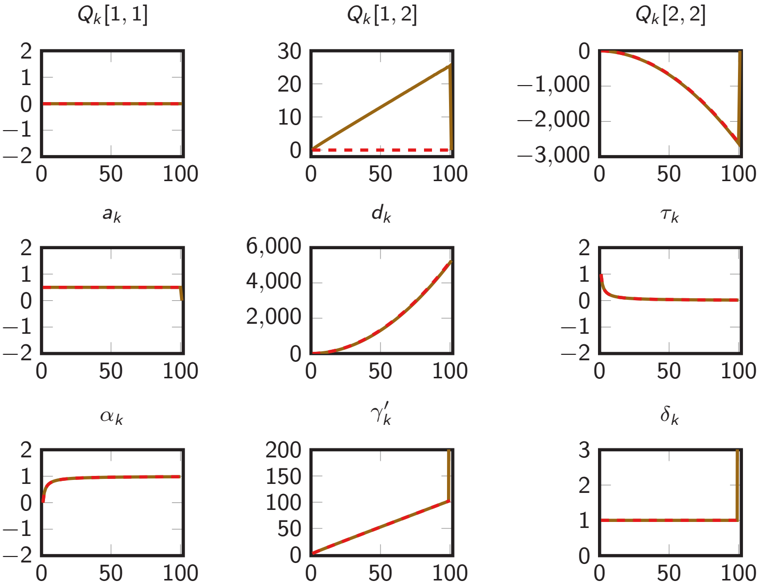
$$\phi_k^f = \begin{pmatrix} y_k - x_\star \\ f'(y_k) \end{pmatrix}^\top [Q_k \otimes I_d] \begin{pmatrix} y_k - x_\star \\ f'(y_k) \end{pmatrix} + a_k \|z_k - x_\star\|^2 + d_k \left(f(y_k) - f_\star\right),$$

we can perform similar steps.

Example for $N = 100$, $L = 1$: numerics (brown),



Example for $N = 100$, $L = 1$: numerics (brown), and analytical solution (red).



Optimized gradient method

Optimized gradient methods (Kim & Fessler, 2016) can be factorized in a similar form

$$\begin{aligned} y_{k+1} &= (1-\tau_k)y_k + \tau_k z_k - \alpha_k f'(y_k), \\ z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_k) - \gamma_k' f'(y_{k+1}), \end{aligned}$$

but current analysis is more involved (not based on potentials).

From numerical inspiration, we get

$$\phi_k^f = d_k''(f(y_k) - f_\star - \tfrac{1}{2L}\left\|f'(y_k)\right\|^2) + \tfrac{L}{2}\|z_k - x_\star\|^2,$$

with $d_k'' \sim k^2$ (more precisely: $d_{k+1}'' = 1 + d_k'' + \sqrt{1 + 2d_k''}$, *red curves on prev. slide*)

Conjugate gradient method

Conjugate gradient method ("ideal version")

$$x_{k+1} = \operatorname{argmin}_x \{f(x) \,:\, x \in x_0 + \operatorname{span}\{f'(x_0), f'(x_1), \ldots, f'(x_k)\}\}.$$

Steps to perform the analysis are slightly trickier (reference at the end), but

- ◇ analysis is exactly the same as that of the optimized gradient method,
- ◇ achieve exactly the lower complexity bound for the class of problems.

Concluding remarks

Overall philosophy:

Concluding remarks

Overall philosophy:

- ◇ numerically obtain best "fixed-horizon" potential-based guarantees,

Concluding remarks

Overall philosophy:

- ◇ numerically obtain best "fixed-horizon" potential-based guarantees,
- ◇ helps designing & benchmarking proofs,

Concluding remarks

Overall philosophy:

- ◇ numerically obtain best "fixed-horizon" potential-based guarantees,
- ◇ helps designing & benchmarking proofs,

More examples?

Concluding remarks

Overall philosophy:

- ◇ numerically obtain best "fixed-horizon" potential-based guarantees,
- ◇ helps designing & benchmarking proofs,

More examples?

- ◇ proximal/projected variants, splitting methods, mirror descent/Bregman gradient, etc.
- ◇ inexact, randomized, and stochastic variants, etc.
- ◇ first attempts on adaptive methods (line searches, Polyak steps),
- ◇ also other classes of functions and problems (nonsmooth, weakly convex, and indicator functions, monotone inclusions, variational inequalities), etc.

… and probably many others!

Concluding remarks

Overall philosophy:

- ◇ numerically obtain best “fixed-horizon” potential-based guarantees,
- ◇ helps designing & benchmarking proofs,

More examples?

- ◇ proximal/projected variants, splitting methods, mirror descent/Bregman gradient, etc.
- ◇ inexact, randomized, and stochastic variants, etc.
- ◇ first attempts on adaptive methods (line searches, Polyak steps),
- ◇ also other classes of functions and problems (nonsmooth, weakly convex, and indicator functions, monotone inclusions, variational inequalities), etc.

... and probably many others!

... and open questions:

Concluding remarks

Overall philosophy:

- ◇ numerically obtain best "fixed-horizon" potential-based guarantees,
- ◇ helps designing & benchmarking proofs,

More examples?

- ◇ proximal/projected variants, splitting methods, mirror descent/Bregman gradient, etc.
- ◇ inexact, randomized, and stochastic variants, etc.
- ◇ first attempts on adaptive methods (line searches, Polyak steps),
- ◇ also other classes of functions and problems (nonsmooth, weakly convex, and indicator functions, monotone inclusions, variational inequalities), etc.

... and probably many others!

... and open questions:

- ◇ beyond Euclidean geometry?
- ◇ Higher-order methods?
- ◇ Adaptive methods (BFGS, nonlinear conjugate gradients)?
- ◇ Beyond worst-case analyses?

A few references

A few references

Shameless advertisement:

- Radu-Alexandru Dragomir, T, Alexandre d'Aspremont, Jérôme Bolte. "Optimal complexity and certification of Bregman first-order methods". Preprint 2019.
- Mathieu Barré, T, Francis Bach. "Principled Analyses and Design of First-Order Methods with Inexact Proximal Operators". Preprint 2020.
- Mathieu Barré, T, Alexandre d'Aspremont. "Complexity Guarantees for Polyak Steps with Momentum". COLT 2020.

A few references

Shameless advertisement:

- Radu-Alexandru Dragomir, T, Alexandre d'Aspremont, Jérôme Bolte. "Optimal complexity and certification of Bregman first-order methods". Preprint 2019.
- Mathieu Barré, T, Francis Bach. "Principled Analyses and Design of First-Order Methods with Inexact Proximal Operators". Preprint 2020.
- Mathieu Barré, T, Alexandre d'Aspremont. "Complexity Guarantees for Polyak Steps with Momentum". COLT 2020.

References more thoroughly treated in the papers. Explicitly mentioned in this presentation:

- Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 1983.
- Amir Beck, Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". SIAM Journal on Imaging Sciences, 2009.
- Yoel Drori, Marc Teboulle. "Performance of first-order methods for smooth convex minimization: a novel approach". Mathematical Programming, 2014.
- Donghwan Kim, Jeffrey Fessler. "Optimized first-order methods for smooth convex minimization". Mathematical Programming, 2016.
- Laurent Lessard, Benjamin Recht, Andrew Packard. "Analysis and design of optimization algorithms via integral quadratic constraints". SIAM Journal on Optimization, 2016.
- Bin Hu, Laurent Lessard. "Dissipativity Theory for Nesterov's Accelerated Method". ICML, 2017.
- Nikhil Bansal, Anupam Gupta. "Potential-function proofs for first-order methods". Theory of Computing, 2019.

Thanks! Questions?

www.di.ens.fr/∼ataylor/

Codes (on Github)

- ⋄ AdrienTaylor/Potential-functions-for-first-order-methods
- ⋄ AdrienTaylor/Performance-Estimation-Toolbox

Thanks! Questions?

www.di.ens.fr/∼ataylor/

Codes (on Github)

- ⋄ AdrienTaylor/Potential-functions-for-first-order-methods
- ⋄ AdrienTaylor/Performance-Estimation-Toolbox

Tutorial (computer-assisted proofs in optimization):

- ⋄ https://francisbach.com/computer-aided-analyses/

Thanks! Questions?

www.di.ens.fr/∼ataylor/

Codes (on Github)

- ◇ AdrienTaylor/Potential-functions-for-first-order-methods
- ◇ AdrienTaylor/Performance-Estimation-Toolbox

Tutorial (computer-assisted proofs in optimization):

- ◇ https://francisbach.com/computer-aided-analyses/

Presentation mainly based on

- ◇ T., Francis Bach. "Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions", 2019.
- ◇ Yoel Drori, T. "Efficient first-order methods for convex minimization: a constructive approach", 2019.
- ◇ T., François Glineur, Julien Hendrickx. "Smooth strongly convex interpolation and exact worst-case performance of first-order methods", 2017.