Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions

Adrien Taylor, Francis Bach

Conference on Learning Theory (COLT) - June 2019

What is this work about?

What is this work about?

Computer-assisted analyses of first-order optimization methods

What is this work 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 work 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), (Bansal & Gupta 2017), (Hu & Lessard 2017), (Wilson, Recht & Jordan 2016), and many others.

Potential functions

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

$$f_\star = \min_{x \in \mathbb{R}^d} f(x)?$$

Potential functions

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

$$f_\star = \min_{x \in \mathbb{R}^d} f(x)?$$

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

$$f_\star = \min_{x \in \mathbb{R}^d} f(x)?$$

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, $x_k \in \mathbb{R}^d$, 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 2017).

Potential functions

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

$$f_\star = \min_{x\in\mathbb{R}^d} f(x)?$$

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, $x_k \in \mathbb{R}^d$, and $k \geq 0$, easy to show $\phi_{k+1}^f \leq \phi_k^f$ with

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

see e.g., (Bansal & Gupta 2017).

Why is that nice? Very simple resulting proof:

Potential functions

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

$$f_\star = \min_{x \in \mathbb{R}^d} f(x)?$$

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, $x_k \in \mathbb{R}^d$, 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 2017).

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

$$f_\star = \min_{x\in\mathbb{R}^d} f(x)?$$

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, $x_k \in \mathbb{R}^d$, 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 2017).

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

$$f_\star = \min_{x\in\mathbb{R}^d} f(x)?$$

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, $x_k \in \mathbb{R}^d$, 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 2017).

Why is that nice? Very simple resulting proof:

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

Potential functions

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

$$f_\star = \min_{x \in \mathbb{R}^d} f(x)?$$

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, $x_k \in \mathbb{R}^d$, 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 2017).

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

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

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, $x_k \in \mathbb{R}^d$, and $d \in \mathbb{N}$:

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

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, $x_k \in \mathbb{R}^d$, and $d \in \mathbb{N}$:

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

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

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, $x_k \in \mathbb{R}^d$, and $d \in \mathbb{N}$:

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

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

Answer:

$$\phi_{k+1}^f \leq \phi_k^f \text{ for all } L\text{-smooth convex } f,\ x_k \in \mathbb{R}^d, \text{ and } d \in \mathbb{N}$$

$$\Leftrightarrow$$

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, $x_k \in \mathbb{R}^d$, and $d \in \mathbb{N}$:

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

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

Answer:

$$\phi^f_{k+1} \leq \phi^f_k \text{ for all } L\text{-smooth convex } f,\ x_k \in \mathbb{R}^d, \text{ and } d \in \mathbb{N}$$

$$\Leftrightarrow$$

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, $x_k \in \mathbb{R}^d$, and $d \in \mathbb{N}$:

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

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

Answer:

$$\phi_{k+1}^f \leq \phi_k^f \text{ for all } L\text{-smooth convex } f,\ x_k \in \mathbb{R}^d, \text{ and } d \in \mathbb{N}$$

$$\Leftrightarrow$$

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

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

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

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

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

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

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

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.

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.
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:

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

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

How does it work for the gradient method?

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

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

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

How does it work for the gradient method?

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

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

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

How does it work for the gradient method?

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

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

$N =$	1	2
$b_N =$	4	9

How does it work for the gradient method?

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

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

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

How does it work for the gradient method?

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

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

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

How does it work for the gradient method?

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

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

$N =$	1	2	3	4	...	100
$b_N =$	4	9	16	25	...	10201

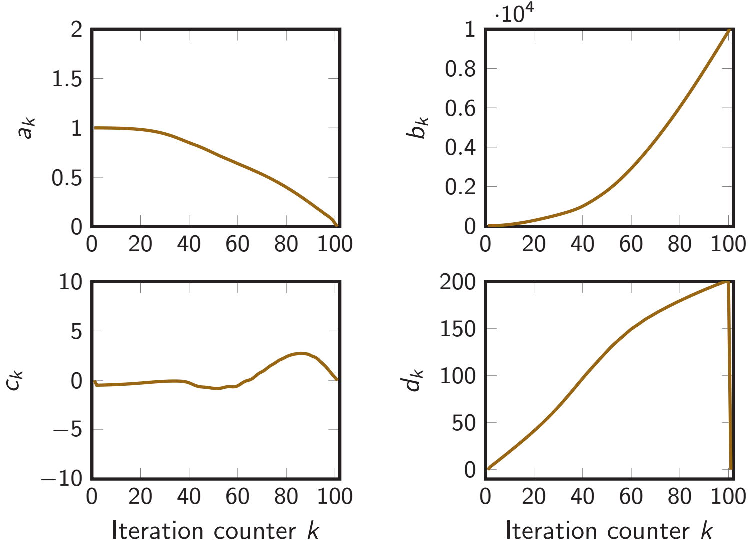
2. Observe the a_k, b_k, c_k, d_k's for some values of N.

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

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

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:

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

$N =$	1	2	3	4	...	100
$b_N =$	4	9	16	25	...	10201

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?

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

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

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

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.

How does it work for the gradient method?

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

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

$N =$	1	2	3	4	...	100
$b_N =$	4	9	16	25	...	10201

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 $\mathcal{V}_k$:

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

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

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,
- *before trying to prove your new crazy first-order method works; give it a try!*

Concluding remarks

Overall philosophy:

- numerically obtain best "fixed-horizon" potential-based guarantees,
- helps designing & benchmarking proofs,
- *before trying to prove your new crazy first-order method works; give it a try!*

More examples in the paper (T. and Bach, 2019):

Concluding remarks

Overall philosophy:

- ◇ numerically obtain best "fixed-horizon" potential-based guarantees,
- ◇ helps designing & benchmarking proofs,
- ◇ *before trying to prove your new crazy first-order method works; give it a try!*

More examples in the paper (T. and Bach, 2019):

- ◇ accelerated variants (also automated parameter selection),
- ◇ proximal variants,
- ◇ stochastic variants (e.g., under bounded variance or over-parametrization),
- ◇ randomized block-coordinate variants,

... and probably many others (but not in the paper)!

Thanks!

Interested? Poster #174

"Stochastic first-order methods:
non-asymptotic and computer-aided analyses
via potential functions"