Computer-aided analyses of first-order methods

(via semidefinite programming)

Adrien Taylor

Cambridge – February 2020

François Glineur
(UCLouvain)

Julien Hendrickx
(UCLouvain)

Etienne de Klerk
(Tilburg & Delft)

Ernest Ryu
(UCLA)

Francis Bach
(Inria/ENS)

Jérôme Bolte
(TSE)

A. d'Aspremont
(CNRS/ENS)

Yoel Drori
(Google)

Mathieu Barré
(Inria/ENS)

A-R. Dragomir
(ENS/TSE)

B. Van Scoy
(W-Madison)

L. Lessard
(W-Madison)

C. Bergeling
(Lund)

P. Giselsson
(Lund)

Take-home messages

Worst-cases are solutions to optimization problems.

Take-home messages

Worst-cases are solutions to optimization problems.

Sometimes, those optimization problems are tractable.

Take-home messages

Worst-cases are solutions to optimization problems.

Sometimes, those optimization problems are tractable.

Often tractable for first-order methods in convex optimization!

Toy example

Performance estimation

Further examples

Toward simpler proofs

Conclusions and discussions

Toy example

Performance estimation

Further examples

Toward simpler proofs

Conclusions and discussions

Analysis of a gradient step

We want to solve

$$\min_{x \in \mathbb{R}^d} f(x)$$

under some assumptions on f.

Analysis of a gradient step

We want to solve

$$\min_{x \in \mathbb{R}^d} f(x)$$

under some assumptions on f.

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

Analysis of a gradient step

We want to solve

$$\min_{x\in\mathbb{R}^d} f(x)$$

under some assumptions on f.

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

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

Analysis of a gradient step

We want to solve

$$\min_{x \in \mathbb{R}^d} f(x)$$

under some assumptions on f.

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

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

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

Convergence rate of a gradient step

Convergence rate of a gradient step

Toy example: Convergence rate: what is the smallest ρ such that?

$$\left\|f'(x_1)\right\| \leq \rho \left\|f'(x_0)\right\|$$

for all $x_0, x_1 \in \mathbb{R}^d$, all f, and $x_1 = x_0 - \gamma f'(x_0)$?

Convergence rate of a gradient step

Toy example: Convergence rate: what is the smallest ρ such that?

$$\left\|f'(x_1)\right\| \leq \rho \left\|f'(x_0)\right\|$$

for all $x_0, x_1 \in \mathbb{R}^d$, all f, and $x_1 = x_0 - \gamma f'(x_0)$?

- Optimization problem to find sharp convergence rate:

$$\begin{aligned} \max_{f, x_0, x_1} \quad & \frac{\|f'(x_1)\|}{\|f'(x_0)\|} \\ \text{subject to} \quad & x_1 \text{ generated by gradient descent from } x_0, \\ & \text{assumptions on } f, \end{aligned}$$

  which has function f as variable.

Assumptions

Assumptions

Nontrivial rates only by assuming something on f.

Assumptions

Nontrivial rates only by assuming something on f.

For example: pick assumptions among the following:

Assumptions

Nontrivial rates only by assuming something on f.

For example: pick assumptions among the following:

- A convex function f is commonly assumed to be (for all $x, y \in \mathbb{R}^d$):
  - μ-strongly convex $\quad f(x) \geq f(y) + \langle \partial f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2,$
  - L-smooth $\quad f(x) \leq f(y) + \langle f'(y), x - y \rangle + \frac{L}{2}\|x - y\|^2.$

Assumptions

Nontrivial rates only by assuming something on f.

For example: pick assumptions among the following:

- A convex function f is commonly assumed to be (for all $x, y \in \mathbb{R}^d$):
  - μ-strongly convex $\quad f(x) \geq f(y) + \langle \partial f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2$,
  - L-smooth $\quad f(x) \leq f(y) + \langle f'(y), x - y \rangle + \frac{L}{2}\|x - y\|^2$.

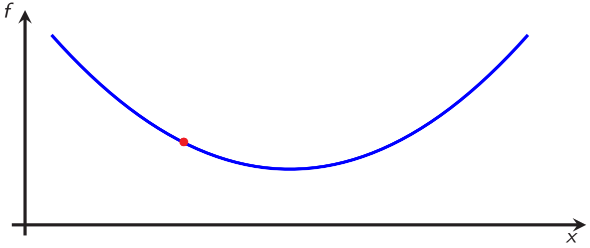
Here, we choose: $f \in \mathcal{F}_{\mu,L}$: class of μ-strongly convex L-smooth functions.

About the assumptions

Consider a differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, f is (μ-strongly) convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$ we have:

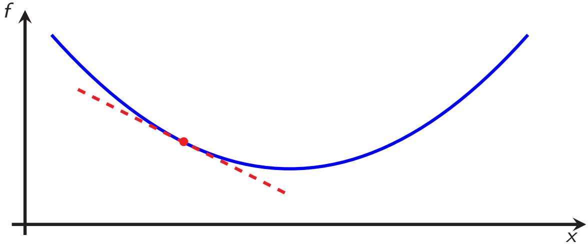
About the assumptions

Consider a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f is (μ-strongly) convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$ we have:



About the assumptions

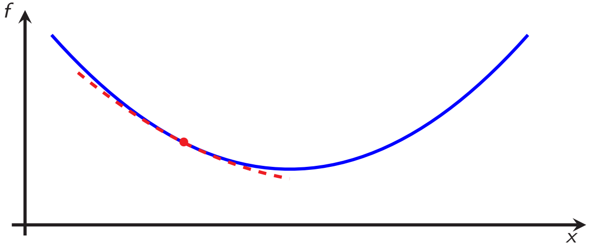
Consider a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f is (μ-strongly) convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$ we have:



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

About the assumptions

Consider a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f is (μ-strongly) convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$ we have:

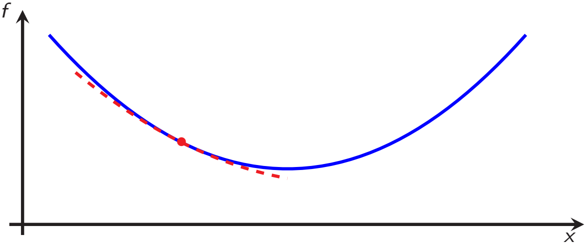


(1) (Convexity) $f(x) \geq f(y) + \langle f'(y), x - y \rangle$,

(1b) (μ-strong convexity) $f(x) \geq f(y) + \langle f'(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2$,

About the assumptions

Consider a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f is (μ-strongly) convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$ we have:



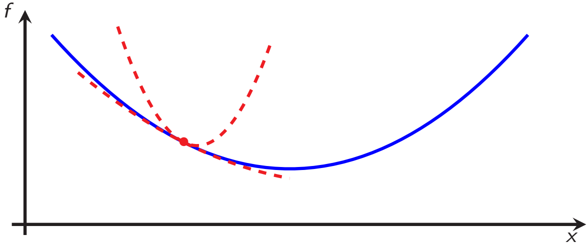
(1) (Convexity) $f(x) \geq f(y) + \langle f'(y), x - y \rangle$,

(1b) (μ-strong convexity) $f(x) \geq f(y) + \langle f'(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2$,

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

About the assumptions

Consider a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f is (μ-strongly) convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$ we have:



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

(1b) (μ-strong convexity) $f(x) \geq f(y) + \langle f'(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2$,

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

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

Convergence rate of a gradient step

Convergence rate of a gradient step

Toy example: Convergence rate: what is the smallest ρ such that?

$$\left\|f'(x_1)\right\| \leq \rho \left\|f'(x_0)\right\|$$

for all $x_0, x_1 \in \mathbb{R}^d$, all f, and $x_1 = x_0 - \gamma f'(x_0)$?

Convergence rate of a gradient step

Toy example: Convergence rate: what is the smallest ρ such that?

$$\left\|f'(x_1)\right\| \leq \rho \left\|f'(x_0)\right\|$$

for all $x_0, x_1 \in \mathbb{R}^d$, all f, and $x_1 = x_0 - \gamma f'(x_0)$?

- Optimization problem to find sharp convergence rate:

$$\begin{aligned} \max_{f, x_0, x_1} \quad & \frac{\|f'(x_1)\|}{\|f'(x_0)\|} \\ \text{subject to} \quad & x_1 \text{ generated by gradient descent from } x_0, \\ & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex.} \end{aligned}$$

  which has function f as variable.

Convergence rate of a gradient step

Toy example: Convergence rate: what is the smallest ρ such that?

$$\left\|f'(x_1)\right\| \leq \rho \left\|f'(x_0)\right\|$$

for all $x_0, x_1 \in \mathbb{R}^d$, all f, and $x_1 = x_0 - \gamma f'(x_0)$?

- Optimization problem to find sharp convergence rate:

$$\max_{f, x_0, x_1} \quad \frac{\|f'(x_1)\|}{\|f'(x_0)\|}$$

$$\text{subject to} \quad x_1 \text{ generated by gradient descent from } x_0,$$

$$f \text{ is } L\text{-smooth and } \mu\text{-strongly convex.}$$

  which has function f as variable.

- Variables: f, x_0, x_1;

Convergence rate of a gradient step

Toy example: Convergence rate: what is the smallest ρ such that?

$$\|f'(x_1)\| \leq \rho \, \|f'(x_0)\|$$

for all $x_0, x_1 \in \mathbb{R}^d$, all f, and $x_1 = x_0 - \gamma f'(x_0)$?

- Optimization problem to find sharp convergence rate:

$$\begin{array}{cl} \displaystyle\max_{f, x_0, x_1} & \dfrac{\|f'(x_1)\|}{\|f'(x_0)\|} \\ \text{subject to} & x_1 \text{ generated by gradient descent from } x_0, \\ & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex.} \end{array}$$

  which has function f as variable.

- Variables: f, x_0, x_1; parameters: μ, L, γ.

Convergence rate of a gradient step

Toy example: Convergence rate: what is the smallest ρ such that?

$$\left\|f'(x_1)\right\| \leq \rho \left\|f'(x_0)\right\|$$

for all $x_0, x_1 \in \mathbb{R}^d$, all f, and $x_1 = x_0 - \gamma f'(x_0)$?

- Optimization problem to find sharp convergence rate:

$$\begin{aligned} \max_{f, x_0, x_1} \quad & \frac{\|f'(x_1)\|}{\|f'(x_0)\|} \\ \text{subject to} \quad & x_1 \text{ generated by gradient descent from } x_0, \\ & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex.} \end{aligned}$$

  which has function f as variable.
- Variables: f, x_0, x_1; parameters: μ, L, γ.
- Optimal value can be found via convex optimization! (3x3 SDP):

$$\max \left\{ \frac{\|f'(x_1)\|^2}{\|f'(x_0)\|^2} \right\} = \max \left\{ (1-\mu\gamma)^2, (1-L\gamma)^2 \right\}$$

From infinite to finite dimensional problems

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

From infinite to finite dimensional problems

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

- How to treat the infinite dimensional variable f?

From infinite to finite dimensional problems

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 $f \in \mathcal{F}_{\mu,L}$?

From infinite to finite dimensional problems

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 $f \in \mathcal{F}_{\mu,L}$?

Idea:

From infinite to finite dimensional problems

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 $f \in \mathcal{F}_{\mu,L}$?

Idea:

- replace f by its discrete version:

$$f_i = f(x_i),\ g_i = f'(x_i) \quad \forall i \in \{0, 1\}\,.$$

From infinite to finite dimensional problems

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 $f \in \mathcal{F}_{\mu,L}$?

Idea:

- replace f by its discrete version:

$$f_i = f(x_i),\ g_i = f'(x_i) \quad \forall i \in \{0,1\}.$$

- Require points (x_i, g_i, f_i) to be interpolable by a function $f \in \mathcal{F}_{\mu,L}$.

From infinite to finite dimensional problems

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 $f \in \mathcal{F}_{\mu,L}$?

Idea:

- replace f by its discrete version:

$$f_i = f(x_i),\ g_i = f'(x_i) \quad \forall i \in \{0,1\}\,.$$

- Require points (x_i, g_i, f_i) to be interpolable by a function $f \in \mathcal{F}_{\mu,L}$.
  The new constraint is:

$$\exists f \in \mathcal{F}_{\mu,L}:\ f_i = f(x_i),\ g_i = f'(x_i), \qquad \forall i \in \{0,1\}\,.$$

Discrete version

Discrete version

- Optimization problem to find sharp convergence rate:

$$
\begin{aligned}
\max_{f, x_0, x_1} \quad & \frac{\|f'(x_1)\|}{\|f'(x_0)\|} \\
\text{subject to} \quad & x_1 \text{ generated by gradient descent from } x_0 \\
& f \text{ is } L\text{-smooth and } \mu\text{-strongly convex.}
\end{aligned}
$$

Discrete version

- Optimization problem to find sharp convergence rate:

$$\begin{array}{ll} \max\limits_{f, x_0, x_1} & \dfrac{\|f'(x_1)\|}{\|f'(x_0)\|} \\ \text{subject to} & x_1 \text{ generated by gradient descent from } x_0 \\ & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex.} \end{array}$$

- Variables: f, x_0, x_1.

Discrete version

- Optimization problem to find sharp convergence rate:

$$\begin{array}{ll} \max\limits_{f, x_0, x_1} & \dfrac{\|f'(x_1)\|}{\|f'(x_0)\|} \\ \text{subject to} & x_1 \text{ generated by gradient descent from } x_0 \\ & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex.} \end{array}$$

- Variables: f, x_0, x_1.
- Discrete version:

$$\begin{array}{ll} \max\limits_{\substack{x_0, x_1, g_0, g_1 \\ f_0, f_1}} & \dfrac{\|g_1\|}{\|g_0\|} \\ \text{subject to} & x_1 = x_0 - \gamma g_0 \\ & \exists f \in \mathcal{F}_{\mu,L} \text{ such that } \begin{cases} f_i = f(x_i) & i = 1, 2 \\ g_i = f'(x_i) & i = 1, 2 \end{cases} \end{array}$$

Discrete version

- Optimization problem to find sharp convergence rate:

$$\begin{array}{ll} \max\limits_{f, x_0, x_1} & \dfrac{\|f'(x_1)\|}{\|f'(x_0)\|} \\ \text{subject to} & x_1 \text{ generated by gradient descent from } x_0 \\ & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex.} \end{array}$$

- Variables: f, x_0, x_1.
- Discrete version:

$$\begin{array}{ll} \max\limits_{\substack{x_0, x_1, g_0, g_1 \\ f_0, f_1}} & \dfrac{\|g_1\|}{\|g_0\|} \\ \text{subject to} & x_1 = x_0 - \gamma g_0 \\ & \exists f \in \mathcal{F}_{\mu,L} \text{ such that } \begin{cases} f_i = f(x_i) & i = 1, 2 \\ g_i = f'(x_i) & i = 1, 2 \end{cases} \end{array}$$

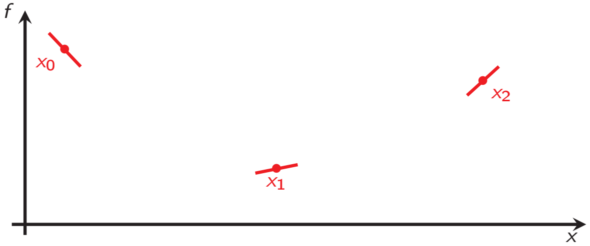
- Variables: x_0, x_1, g_0, g_1, f_0, f_1.

Smooth strongly convex interpolation

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

Smooth strongly convex interpolation

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

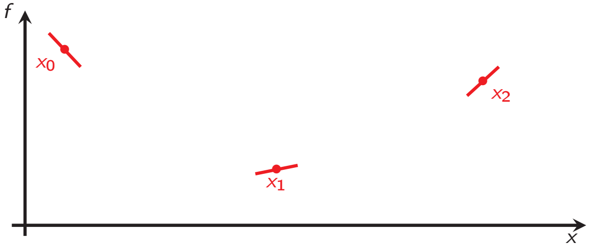


? Possible to find $f \in \mathcal{F}_{\mu,L}$ such that

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

Smooth strongly convex interpolation

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



? Possible to find $f \in \mathcal{F}_{\mu,L}$ such that

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

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

$$f_i \geq f_j + \langle g_j, x_i - x_j \rangle + \frac{1}{2L}\left\|g_i - g_j\right\|^2 + \frac{\mu}{2(1-\mu/L)}\left\|x_i - x_j - \frac{1}{L}(g_i - g_j)\right\|^2.$$

Replace constraints

Replace constraints

◇ Interpolation conditions allow removing red constraints

$$
\begin{aligned}
\max_{\substack{x_0, x_1, g_0, g_1 \\ f_0, f_1}} \quad & \frac{\|g_1\|}{\|g_0\|} \\
\text{subject to} \quad & x_1 = x_0 - \gamma g_0, \\
& \exists f \in \mathcal{F}_{\mu,L} \text{ such that } \begin{cases} f_i = f(x_i) & i = 1,2 \\ g_i = f'(x_i) & i = 1,2 \end{cases}
\end{aligned}
$$

Replace constraints

◇ Interpolation conditions allow removing red constraints

$$
\begin{aligned}
\max_{\substack{x_0, x_1, g_0, g_1\\ f_0, f_1}} \quad & \frac{\|g_1\|}{\|g_0\|} \\
\text{subject to} \quad & x_1 = x_0 - \gamma g_0, \\
& \exists f \in \mathcal{F}_{\mu,L} \text{ such that } \begin{cases} f_i = f(x_i) & i = 1,2 \\ g_i = f'(x_i) & i = 1,2 \end{cases}
\end{aligned}
$$

◇ replacing them by

$$
\begin{aligned}
f_1 &\geq f_0 + \langle g_0, x_1 - x_0 \rangle + \tfrac{1}{2L}\|g_1 - g_0\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_1 - x_0 - \tfrac{1}{L}(g_1 - g_0)\right\|^2 \\
f_0 &\geq f_1 + \langle g_1, x_0 - x_1 \rangle + \tfrac{1}{2L}\|g_0 - g_1\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(g_0 - g_1)\right\|^2.
\end{aligned}
$$

Replace constraints

◇ Interpolation conditions allow removing red constraints

$$
\begin{aligned}
\max_{\substack{x_0, x_1, g_0, g_1 \\ f_0, f_1}} \quad & \frac{\|g_1\|}{\|g_0\|} \\
\text{subject to} \quad & x_1 = x_0 - \gamma g_0, \\
& \exists f \in \mathcal{F}_{\mu,L} \text{ such that } \begin{cases} f_i = f(x_i) & i = 1,2 \\ g_i = f'(x_i) & i = 1,2 \end{cases}
\end{aligned}
$$

◇ replacing them by

$$
\begin{aligned}
f_1 &\geq f_0 + \langle g_0, x_1 - x_0 \rangle + \tfrac{1}{2L}\|g_1 - g_0\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_1 - x_0 - \tfrac{1}{L}(g_1 - g_0)\right\|^2 \\
f_0 &\geq f_1 + \langle g_1, x_0 - x_1 \rangle + \tfrac{1}{2L}\|g_0 - g_1\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(g_0 - g_1)\right\|^2.
\end{aligned}
$$

◇ Same optimal value (no relaxation); but still non-convex quadratic problem.

Reformulations (cont'd)

Reformulations (cont'd)

◇ Equivalent problem: replace red constraints

$$
\begin{aligned}
\max_{\substack{x_0, x_1, g_0, g_1 \\ f_0, f_1}} \quad & \frac{\|g_1\|}{\|g_0\|} \\
\text{subject to} \quad & x_1 = x_0 - \gamma g_0, \\
& f_1 \geq f_0 + \langle g_0, x_1 - x_0 \rangle + \tfrac{1}{2L}\|g_1 - g_0\|^2 \\
& \qquad + \tfrac{\mu}{2(1-\mu/L)}\left\|x_1 - x_0 - \tfrac{1}{L}(g_1 - g_0)\right\|^2 \\
& f_0 \geq f_1 + \langle g_1, x_0 - x_1 \rangle + \tfrac{1}{2L}\|g_0 - g_1\|^2 \\
& \qquad + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(g_0 - g_1)\right\|^2.
\end{aligned}
$$

Reformulations (cont'd)

◇ Equivalent problem: replace red constraints

$$
\begin{aligned}
\max_{\substack{x_0, x_1, g_0, g_1 \\ f_0, f_1}} \quad & \frac{\|g_1\|}{\|g_0\|} \\
\text{subject to} \quad & x_1 = x_0 - \gamma g_0, \\
& f_1 \geq f_0 + \langle g_0, x_1 - x_0 \rangle + \tfrac{1}{2L}\|g_1 - g_0\|^2 \\
& \qquad + \tfrac{\mu}{2(1-\mu/L)}\left\|x_1 - x_0 - \tfrac{1}{L}(g_1 - g_0)\right\|^2 \\
& f_0 \geq f_1 + \langle g_1, x_0 - x_1 \rangle + \tfrac{1}{2L}\|g_0 - g_1\|^2 \\
& \qquad + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(g_0 - g_1)\right\|^2.
\end{aligned}
$$

◇ by (substitute $x_1 = x_0 - \gamma g_0$):

$$
\begin{aligned}
f_1 &\geq f_0 - \gamma\|g_0\|^2 + \tfrac{1}{2L}\|g_1 - g_0\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|-\gamma g_0 - \tfrac{1}{L}(g_1 - g_0)\right\|^2 \\
f_0 &\geq f_1 + \gamma\langle g_1, g_0 \rangle + \tfrac{1}{2L}\|g_0 - g_1\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|\gamma g_0 - \tfrac{1}{L}(g_0 - g_1)\right\|^2.
\end{aligned}
$$

Semidefinite lifting

Semidefinite lifting

◇ All elements are quadratic in (x_0, g_0, g_1), and linear in (f_0, f_1):

$$\max_{\substack{x_0, x_1, g_0, g_1\\ f_0, f_1}} \quad \frac{\|g_1\|}{\|g_0\|}$$

$$\text{subject to} \quad f_1 \geq f_0 - \gamma\|g_0\|^2 + \frac{1}{2L}\|g_1 - g_0\|^2 + \frac{\mu}{2(1-\mu/L)}\left\|-\gamma g_0 - \frac{1}{L}(g_1 - g_0)\right\|^2$$

$$f_0 \geq f_1 + \gamma\langle g_1, g_0\rangle + \frac{1}{2L}\|g_0 - g_1\|^2 + \frac{\mu}{2(1-\mu/L)}\left\|\gamma g_0 - \frac{1}{L}(g_0 - g_1)\right\|^2.$$

Semidefinite lifting

- All elements are quadratic in (x_0, g_0, g_1), and linear in (f_0, f_1):

$$\max_{\substack{x_0, x_1, g_0, g_1 \\ f_0, f_1}} \quad \frac{\|g_1\|}{\|g_0\|}$$

$$\text{subject to} \quad f_1 \geq f_0 - \gamma\|g_0\|^2 + \tfrac{1}{2L}\|g_1 - g_0\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|-\gamma g_0 - \tfrac{1}{L}(g_1 - g_0)\right\|^2$$

$$f_0 \geq f_1 + \gamma\langle g_1, g_0\rangle + \tfrac{1}{2L}\|g_0 - g_1\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|\gamma g_0 - \tfrac{1}{L}(g_0 - g_1)\right\|^2.$$

- They can therefore be represented with a Gram matrix G and a vector F, with

$$G = \begin{bmatrix} \|x_0\|^2 & \langle x_0, g_0\rangle & \langle x_0, g_1\rangle \\ \langle x_0, g_0\rangle & \|g_0\|^2 & \langle g_0, g_1\rangle \\ \langle x_0, g_1\rangle & \langle g_0, g_1\rangle & \|g_1\|^2 \end{bmatrix}, \quad F = \begin{bmatrix} f_0 & f_1 \end{bmatrix},$$

where $G \succeq 0$ by construction

Semidefinite lifting

- All elements are quadratic in (x_0, g_0, g_1), and linear in (f_0, f_1):

$$\max_{\substack{x_0, x_1, g_0, g_1\\ f_0, f_1}} \quad \frac{\|g_1\|}{\|g_0\|}$$

$$\text{subject to} \quad f_1 \geq f_0 - \gamma\|g_0\|^2 + \tfrac{1}{2L}\|g_1 - g_0\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|-\gamma g_0 - \tfrac{1}{L}(g_1 - g_0)\right\|^2$$

$$f_0 \geq f_1 + \gamma\langle g_1, g_0\rangle + \tfrac{1}{2L}\|g_0 - g_1\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|\gamma g_0 - \tfrac{1}{L}(g_0 - g_1)\right\|^2.$$

- They can therefore be represented with a Gram matrix G and a vector F, with

$$G = \begin{bmatrix} \|x_0\|^2 & \langle x_0, g_0\rangle & \langle x_0, g_1\rangle \\ \langle x_0, g_0\rangle & \|g_0\|^2 & \langle g_0, g_1\rangle \\ \langle x_0, g_1\rangle & \langle g_0, g_1\rangle & \|g_1\|^2 \end{bmatrix}, \quad F = \begin{bmatrix} f_0 & f_1 \end{bmatrix},$$

where $G \succeq 0$ by construction, and reformulate to:

$$\begin{aligned} \max_{G,\,F} \quad & \frac{b_o^\top F + \mathrm{Tr}(A_o G)}{b_s^\top F + \mathrm{Tr}(A_s G)} \\ \text{subject to} \quad & b_1^\top F + \mathrm{Tr}(A_1 G) \geq 0 \\ & b_2^\top F + \mathrm{Tr}(A_2 G) \geq 0 \\ & G \succeq 0. \end{aligned}$$

with appropriate $A_o, A_s, A_1, A_2, b_o, b_s, b_1, b_2$ for picking elements in G and F.

Semidefinite lifting

- All elements are quadratic in (x_0, g_0, g_1), and linear in (f_0, f_1):

$$
\begin{aligned}
\max_{\substack{x_0,x_1,g_0,g_1\\ f_0,f_1}} \quad & \frac{\|g_1\|}{\|g_0\|} \\
\text{subject to} \quad & f_1 \geq f_0 - \gamma\|g_0\|^2 + \tfrac{1}{2L}\|g_1 - g_0\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|-\gamma g_0 - \tfrac{1}{L}(g_1 - g_0)\right\|^2 \\
& f_0 \geq f_1 + \gamma\langle g_1, g_0\rangle + \tfrac{1}{2L}\|g_0 - g_1\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|\gamma g_0 - \tfrac{1}{L}(g_0 - g_1)\right\|^2.
\end{aligned}
$$

- They can therefore be represented with a Gram matrix G and a vector F, with

$$
G = \begin{bmatrix} \|x_0\|^2 & \langle x_0, g_0\rangle & \langle x_0, g_1\rangle \\ \langle x_0, g_0\rangle & \|g_0\|^2 & \langle g_0, g_1\rangle \\ \langle x_0, g_1\rangle & \langle g_0, g_1\rangle & \|g_1\|^2 \end{bmatrix}, \quad F = \begin{bmatrix} f_0 & f_1 \end{bmatrix},
$$

where $G \succeq 0$ by construction, and reformulate to:

$$
\begin{aligned}
\max_{G,\,F} \quad & \frac{b_o^\top F + \mathrm{Tr}(A_o G)}{b_s^\top F + \mathrm{Tr}(A_s G)} \\
\text{subject to} \quad & b_1^\top F + \mathrm{Tr}(A_1 G) \geq 0 \\
& b_2^\top F + \mathrm{Tr}(A_2 G) \geq 0 \\
& G \succeq 0.
\end{aligned}
$$

with appropriate $A_o, A_s, A_1, A_2, b_o, b_s, b_1, b_2$ for picking elements in G and F.

- Note: assuming $x_0, g_0, g_1 \in \mathbb{R}^d$ with $d \geq 3$, same optimal cost!

Last part in convexification

Last part in convexification

- Constraints are positively homogeneous of deg. 1 and the cost is constant under scaling of G and F

$$
\begin{aligned}
\max_{G,\,F} \quad & \frac{b_o^\top F + \mathrm{Tr}(A_o G)}{b_s^\top F + \mathrm{Tr}(A_s G)} \\
\text{subject to} \quad & b_1^\top F + \mathrm{Tr}(A_1 G) \geq 0 \\
& b_2^\top F + \mathrm{Tr}(A_2 G) \geq 0 \\
& G \succeq 0.
\end{aligned}
$$

Last part in convexification

- Constraints are positively homogeneous of deg. 1 and the cost is constant under scaling of G and F

$$
\begin{array}{rl}
\displaystyle\max_{G,\,F} & \dfrac{b_o^\top F + \mathrm{Tr}(A_o G)}{b_s^\top F + \mathrm{Tr}(A_s G)} \\
\text{subject to} & b_1^\top F + \mathrm{Tr}(A_1 G) \geq 0 \\
 & b_2^\top F + \mathrm{Tr}(A_2 G) \geq 0 \\
 & G \succeq 0.
\end{array}
$$

- Therefore an equivalent *convex* problem is

$$
\begin{array}{rl}
\displaystyle\max_{G,\,F} & b_o^\top F + \mathrm{Tr}(A_o G) \\
\text{subject to} & b_1^\top F + \mathrm{Tr}(A_1 G) \geq 0 \\
 & b_2^\top F + \mathrm{Tr}(A_2 G) \geq 0 \\
 & b_s^\top F + \mathrm{Tr}(A_s G) = 1 \\
 & G \succeq 0.
\end{array}
$$

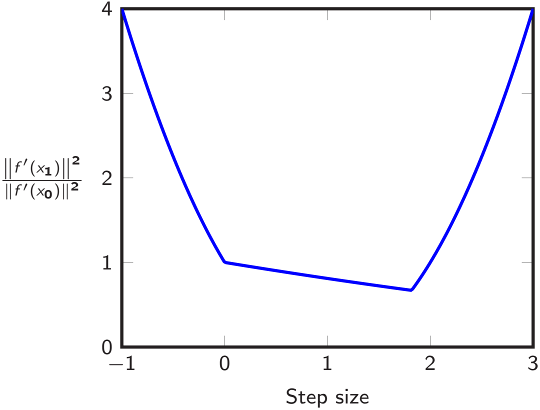
which is a 3x3 semidefinite program.

Solving the SDP...

Fix $L = 1$, $\mu = .1$ and solve the SDP for a few values of γ.

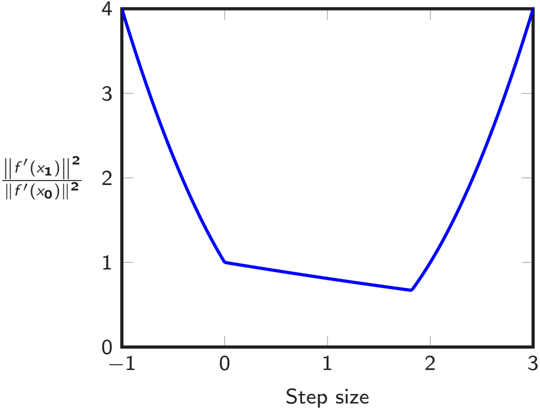
Solving the SDP...

Fix $L = 1$, $\mu = .1$ and solve the SDP for a few values of γ.



Solving the SDP...

Fix $L = 1$, $\mu = .1$ and solve the SDP for a few values of γ.



Observation: it matches $\max\{(1-\gamma L)^2, (1-\gamma\mu)^2\}$—convergence for $\gamma \in (0, 2/L)$.

Dual problem

Dual problem

- Introduce dual variables τ, λ_1 and λ_2

$$
\begin{array}{lll}
\max\limits_{G,\,F} & b_o^\top F + \mathrm{Tr}(A_o G) & \\
\text{subject to} & b_1^\top F + \mathrm{Tr}(A_1 G) \geq 0 & :\ \lambda_1 \\
& b_2^\top F + \mathrm{Tr}(A_2 G) \geq 0 & :\ \lambda_2 \\
& b_s^\top F + \mathrm{Tr}(A_s G) = 1 & :\ \tau \\
& G \succeq 0. &
\end{array}
$$

Dual problem

- Introduce dual variables τ, λ_1 and λ_2

$$\begin{array}{lll} \max\limits_{G,\,F} & b_o^\top F + \mathrm{Tr}(A_o G) & \\ \text{subject to} & b_1^\top F + \mathrm{Tr}(A_1 G) \geq 0 & : \lambda_1 \\ & b_2^\top F + \mathrm{Tr}(A_2 G) \geq 0 & : \lambda_2 \\ & b_s^\top F + \mathrm{Tr}(A_s G) = 1 & : \tau \\ & G \succeq 0. & \end{array}$$

- Dual problem becomes

$$\begin{array}{ll} \underset{\tau,\lambda_1,\lambda_2}{\text{minimize}} & \tau \\ \text{subject to} & \lambda_i \geq 0 \\ & S = A_o + \sum_{i=1}^2 \lambda_i A_i - \tau A_s \preceq 0 \\ & 0 = b_o + \sum_{i=1}^2 \lambda_i b_i - \tau b_s. \end{array}$$

Dual problem

- Introduce dual variables τ, λ_1 and λ_2

$$
\begin{array}{lll}
\max\limits_{G,\,F} & b_o^\top F + \mathrm{Tr}(A_o G) & \\
\text{subject to} & b_1^\top F + \mathrm{Tr}(A_1 G) \geq 0 & : \lambda_1 \\
 & b_2^\top F + \mathrm{Tr}(A_2 G) \geq 0 & : \lambda_2 \\
 & b_s^\top F + \mathrm{Tr}(A_s G) = 1 & : \tau \\
 & G \succeq 0. &
\end{array}
$$

- Dual problem becomes

$$
\begin{array}{ll}
\underset{\tau,\lambda_1,\lambda_2}{\text{minimize}} & \tau \\
\text{subject to} & \lambda_i \geq 0 \\
 & S = A_o + \sum_{i=1}^{2} \lambda_i A_i - \tau A_s \preceq 0 \\
 & 0 = b_o + \sum_{i=1}^{2} \lambda_i b_i - \tau b_s.
\end{array}
$$

- In this example:

$$
S = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\frac{\lambda_1(\gamma\mu-1)(\gamma L-1)}{L-\mu} - \tau & -\frac{\lambda_1(\gamma(\mu+L)-2)}{2(L-\mu)} \\ 0 & -\frac{\lambda_1(\gamma(\mu+L)-2)}{2(L-\mu)} & 1 - \frac{\lambda_1}{L-\mu} \end{bmatrix}
$$

$$
0 = \lambda_1 - \lambda_2.
$$

Dual problem

- ◇ Introduce dual variables τ, λ_1 and λ_2

$$\begin{array}{lll} \max\limits_{G,\,F} & b_o^\top F + \mathrm{Tr}(A_o G) & \\ \text{subject to} & b_1^\top F + \mathrm{Tr}(A_1 G) \geq 0 & : \lambda_1 \\ & b_2^\top F + \mathrm{Tr}(A_2 G) \geq 0 & : \lambda_2 \\ & b_s^\top F + \mathrm{Tr}(A_s G) = 1 & : \tau \\ & G \succeq 0. & \end{array}$$

- ◇ Dual problem becomes

$$\begin{array}{ll} \underset{\tau,\lambda_1,\lambda_2}{\text{minimize}} & \tau \\ \text{subject to} & \lambda_i \geq 0 \\ & S = A_o + \sum_{i=1}^2 \lambda_i A_i - \tau A_s \preceq 0 \\ & 0 = b_o + \sum_{i=1}^2 \lambda_i b_i - \tau b_s. \end{array}$$

- ◇ In this example:

$$S = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\frac{\lambda_1(\gamma\mu-1)(\gamma L-1)}{L-\mu} - \tau & -\frac{\lambda_1(\gamma(\mu+L)-2)}{2(L-\mu)} \\ 0 & -\frac{\lambda_1(\gamma(\mu+L)-2)}{2(L-\mu)} & 1 - \frac{\lambda_1}{L-\mu} \end{bmatrix}$$

$$0 = \lambda_1 - \lambda_2.$$

- ◇ Strong duality holds (existence of a Slater point): $\mathrm{rank}(G) + \mathrm{rank}(S) \leq 3$.

Remarks

A few notes:

Remarks

A few notes:

- Dual interpretation: find smallest convergence rate that can be proved by a linear combination of interpolation inequalities.

Remarks

A few notes:

- ◇ Dual interpretation: find smallest convergence rate that can be proved by a linear combination of interpolation inequalities.
- ◇ Consequence of strong duality: in such settings, any (dimension-independent) convergence rate can be proved by a linear combination of interpolation inequalities.

Remarks

A few notes:

- Dual interpretation: find smallest convergence rate that can be proved by a linear combination of interpolation inequalities.
- Consequence of strong duality: in such settings, any (dimension-independent) convergence rate can be proved by a linear combination of interpolation inequalities.
- The methodology offers 3 ways to proceed:
  - play with primal formulation,
  - play with primal-dual saddle-point formulation,
  - play with dual formulation.

Remarks

A few notes:

- ◇ Dual interpretation: find smallest convergence rate that can be proved by a linear combination of interpolation inequalities.
- ◇ Consequence of strong duality: in such settings, any (dimension-independent) convergence rate can be proved by a linear combination of interpolation inequalities.
- ◇ The methodology offers 3 ways to proceed:
  - play with primal formulation,
  - play with primal-dual saddle-point formulation,
  - play with dual formulation.
- ◇ Any dual feasible point can be translated into a "traditional" (SDP-less) proof.

Remarks

A few notes:

- ◇ Dual interpretation: find smallest convergence rate that can be proved by a linear combination of interpolation inequalities.
- ◇ Consequence of strong duality: in such settings, any (dimension-independent) convergence rate can be proved by a linear combination of interpolation inequalities.
- ◇ The methodology offers 3 ways to proceed:
  - play with primal formulation,
  - play with primal-dual saddle-point formulation,
  - play with dual formulation.
- ◇ Any dual feasible point can be translated into a "traditional" (SDP-less) proof.
- ◇ Standard tricks apply, e.g., trace minimization for promoting low-rank solutions.

Dual problem: find a proof

Gradient with $\gamma = \frac{1}{L}$: combine corresponding inequalities

Dual problem: find a proof

Gradient with $\gamma = \frac{1}{L}$: combine corresponding inequalities

$$\begin{aligned}
f_0 \geq f_1 \quad & +\langle f'(x_1), x_0 - x_1\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
& + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 && : \lambda_1 \\
f_1 \geq f_0 \quad & +\langle f'(x_0), x_1 - x_0\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
& + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 && : \lambda_2
\end{aligned}$$

Dual problem: find a proof

Gradient with $\gamma = \frac{1}{L}$: combine corresponding inequalities

$$
\begin{aligned}
f_0 \geq f_1 \quad & +\langle f'(x_1), x_0 - x_1\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
& + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 \qquad : \lambda_1 \\
f_1 \geq f_0 \quad & +\langle f'(x_0), x_1 - x_0\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
& + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 \qquad : \lambda_2
\end{aligned}
$$

Weighted sum with $\lambda_1, \lambda_2 \geq 0$ can be reformulated as

Dual problem: find a proof

Gradient with $\gamma = \frac{1}{L}$: combine corresponding inequalities

$$\begin{aligned}
f_0 \geq f_1 \quad &+\langle f'(x_1), x_0 - x_1\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
&+\tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 && : \lambda_1 = \tfrac{2}{\gamma}(1-\mu\gamma) \\
f_1 \geq f_0 \quad &+\langle f'(x_0), x_1 - x_0\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
&+\tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 && : \lambda_2 = \tfrac{2}{\gamma}(1-\mu\gamma)
\end{aligned}$$

Weighted sum with $\lambda_1, \lambda_2 \geq 0$ can be reformulated as

Dual problem: find a proof

Gradient with $\gamma = \frac{1}{L}$: combine corresponding inequalities

$$\begin{aligned} f_0 \geq f_1 \quad & + \langle f'(x_1), x_0 - x_1 \rangle + \tfrac{1}{2L} \|f'(x_0) - f'(x_1)\|^2 \\ & + \tfrac{\mu}{2(1-\mu/L)} \left\| x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1)) \right\|^2 \end{aligned} \qquad : \lambda_1 = \tfrac{2}{\gamma}(1 - \mu\gamma)$$

$$\begin{aligned} f_1 \geq f_0 \quad & + \langle f'(x_0), x_1 - x_0 \rangle + \tfrac{1}{2L} \|f'(x_0) - f'(x_1)\|^2 \\ & + \tfrac{\mu}{2(1-\mu/L)} \left\| x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1)) \right\|^2 \end{aligned} \qquad : \lambda_2 = \tfrac{2}{\gamma}(1 - \mu\gamma)$$

Weighted sum with $\lambda_1, \lambda_2 \geq 0$ can be reformulated as

$$(1 - \gamma\mu)^2 \left\| f'(x_0) \right\|^2 \geq \left\| f'(x_1) \right\|^2 + \underbrace{\frac{2 - \gamma(L + \mu)}{\gamma(L - \mu)} \left\| (1 - \mu\gamma) f'(x_0) - f'(x_1) \right\|^2}_{},$$

Dual problem: find a proof

Gradient with $\gamma = \frac{1}{L}$: combine corresponding inequalities

$$
\begin{aligned}
f_0 \geq f_1 \quad & +\langle f'(x_1), x_0 - x_1\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
& + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 \qquad : \lambda_1 = \tfrac{2}{\gamma}(1-\mu\gamma) \\
f_1 \geq f_0 \quad & +\langle f'(x_0), x_1 - x_0\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
& + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 \qquad : \lambda_2 = \tfrac{2}{\gamma}(1-\mu\gamma)
\end{aligned}
$$

Weighted sum with $\lambda_1, \lambda_2 \geq 0$ can be reformulated as

$$
(1-\gamma\mu)^2 \left\|f'(x_0)\right\|^2 \geq \left\|f'(x_1)\right\|^2 + \underbrace{\frac{2-\gamma(L+\mu)}{\gamma(L-\mu)}\left\|(1-\mu\gamma)f'(x_0) - f'(x_1)\right\|^2}_{\geq 0},
$$

Dual problem: find a proof

Gradient with $\gamma = \frac{1}{L}$: combine corresponding inequalities

$$
\begin{aligned}
f_0 \geq f_1 \quad & + \langle f'(x_1), x_0 - x_1 \rangle + \tfrac{1}{2L} \|f'(x_0) - f'(x_1)\|^2 \\
& + \tfrac{\mu}{2(1-\mu/L)} \left\| x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1)) \right\|^2 \qquad : \lambda_1 = \tfrac{2}{\gamma}(1-\mu\gamma) \\
f_1 \geq f_0 \quad & + \langle f'(x_0), x_1 - x_0 \rangle + \tfrac{1}{2L} \|f'(x_0) - f'(x_1)\|^2 \\
& + \tfrac{\mu}{2(1-\mu/L)} \left\| x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1)) \right\|^2 \qquad : \lambda_2 = \tfrac{2}{\gamma}(1-\mu\gamma)
\end{aligned}
$$

Weighted sum with $\lambda_1, \lambda_2 \geq 0$ can be reformulated as

$$
\begin{aligned}
(1-\gamma\mu)^2 \left\| f'(x_0) \right\|^2 & \geq \left\| f'(x_1) \right\|^2 + \underbrace{\frac{2-\gamma(L+\mu)}{\gamma(L-\mu)} \left\| (1-\mu\gamma) f'(x_0) - f'(x_1) \right\|^2}_{\geq 0}, \\
& \geq \left\| f'(x_1) \right\|^2,
\end{aligned}
$$

Dual problem: find a proof

Gradient with $\gamma = \frac{1}{L}$: combine corresponding inequalities

$$\begin{aligned} f_0 \geq f_1 \quad & +\langle f'(x_1), x_0 - x_1\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\ & + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 \end{aligned} \qquad : \lambda_1 = \frac{2}{\gamma}(1-\mu\gamma)$$

$$\begin{aligned} f_1 \geq f_0 \quad & +\langle f'(x_0), x_1 - x_0\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\ & + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 \end{aligned} \qquad : \lambda_2 = \frac{2}{\gamma}(1-\mu\gamma)$$

Weighted sum with $\lambda_1, \lambda_2 \geq 0$ can be reformulated as

$$\begin{aligned} (1-\gamma\mu)^2 \left\|f'(x_0)\right\|^2 & \geq \left\|f'(x_1)\right\|^2 + \underbrace{\frac{2-\gamma(L+\mu)}{\gamma(L-\mu)}\left\|(1-\mu\gamma)f'(x_0) - f'(x_1)\right\|^2}_{\geq 0}, \\ & \geq \left\|f'(x_1)\right\|^2, \end{aligned}$$

leading to $\|f'(x_1)\|^2 \leq (1-\frac{\mu}{L})^2\|f'(x_0)\|^2$

Dual problem: find a proof

Gradient with $\gamma = \frac{1}{L}$: combine corresponding inequalities

$$\begin{aligned}
f_0 \geq f_1 \quad & +\langle f'(x_1), x_0 - x_1\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
& + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 && : \lambda_1 = \tfrac{2}{\gamma}(1-\mu\gamma) \\
f_1 \geq f_0 \quad & +\langle f'(x_0), x_1 - x_0\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
& + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 && : \lambda_2 = \tfrac{2}{\gamma}(1-\mu\gamma)
\end{aligned}$$

Weighted sum with $\lambda_1, \lambda_2 \geq 0$ can be reformulated as

$$\begin{aligned}
(1-\gamma\mu)^2 \left\|f'(x_0)\right\|^2 & \geq \left\|f'(x_1)\right\|^2 + \underbrace{\frac{2-\gamma(L+\mu)}{\gamma(L-\mu)}\left\|(1-\mu\gamma)f'(x_0) - f'(x_1)\right\|^2}_{\geq 0,\ \text{or } = 0 \text{ when worst-case is achieved}}, \\
& \geq \left\|f'(x_1)\right\|^2,
\end{aligned}$$

leading to $\|f'(x_1)\|^2 \leq (1-\frac{\mu}{L})^2\|f'(x_0)\|^2$

Dual problem: find a proof

Gradient with $\gamma = \frac{1}{L}$: combine corresponding inequalities

$$
\begin{aligned}
f_0 \geq f_1 \quad &+\langle f'(x_1), x_0 - x_1\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
&+\tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 && : \lambda_1 = \tfrac{2}{\gamma}(1-\mu\gamma) \\
f_1 \geq f_0 \quad &+\langle f'(x_0), x_1 - x_0\rangle + \tfrac{1}{2L}\|f'(x_0) - f'(x_1)\|^2 \\
&+\tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_1 - \tfrac{1}{L}(f'(x_0) - f'(x_1))\right\|^2 && : \lambda_2 = \tfrac{2}{\gamma}(1-\mu\gamma)
\end{aligned}
$$

Weighted sum with $\lambda_1, \lambda_2 \geq 0$ can be reformulated as

$$
\begin{aligned}
(1-\gamma\mu)^2\left\|f'(x_0)\right\|^2 &\geq \left\|f'(x_1)\right\|^2 + \underbrace{\frac{2-\gamma(L+\mu)}{\gamma(L-\mu)}\left\|(1-\mu\gamma)f'(x_0) - f'(x_1)\right\|^2}_{\geq 0,\ \text{or } = 0 \text{ when worst-case is achieved}}, \\
&\geq \left\|f'(x_1)\right\|^2,
\end{aligned}
$$

leading to $\|f'(x_1)\|^2 \leq (1-\frac{\mu}{L})^2\|f'(x_0)\|^2$ (tight).

PEP genealogy ("my humble, biased, view on...")

Base methodological developments:

PEP genealogy ("my humble, biased, view on...")

Base methodological developments:

'14 Drori and Teboulle (MP): upper bounds on worst-case behaviors of FO methods via SDP. Problems scale with number of iterations ($N \times N$ SDP matrices).

PEP genealogy ("my humble, biased, view on...")

Base methodological developments:

'14 Drori and Teboulle (MP): upper bounds on worst-case behaviors of FO methods via SDP. Problems scale with number of iterations ($N \times N$ SDP matrices).

'16 Kim and Fessler (MP): design of an optimized method for smooth convex minimization, using SDPs.

PEP genealogy ("my humble, biased, view on...")

Base methodological developments:

'14 Drori and Teboulle (MP): upper bounds on worst-case behaviors of FO methods via SDP. Problems scale with number of iterations ($N{\times}N$ SDP matrices).

'16 Kim and Fessler (MP): design of an optimized method for smooth convex minimization, using SDPs.

'16 Lessard, Recht, Packard (SIOPT): smaller SDPs for linear convergence, via integral quadratic constraints ("IQCs"). Essentially Lyapunov functions.

In this presentation:

PEP genealogy ("my humble, biased, view on...")

Base methodological developments:

'14 Drori and Teboulle (MP): upper bounds on worst-case behaviors of FO methods via SDP. Problems scale with number of iterations ($N \times N$ SDP matrices).

'16 Kim and Fessler (MP): design of an optimized method for smooth convex minimization, using SDPs.

'16 Lessard, Recht, Packard (SIOPT): smaller SDPs for linear convergence, via integral quadratic constraints ("IQCs"). Essentially Lyapunov functions.

In this presentation:

'17 T, Hendrickx and Glineur (MP): tightness and primal/dual interpretations of the certificates. (essentially previous slides)

PEP genealogy (“my humble, biased, view on...”)

Base methodological developments:

- ’14 Drori and Teboulle (MP): upper bounds on worst-case behaviors of FO methods via SDP. Problems scale with number of iterations ($N \times N$ SDP matrices).
- ’16 Kim and Fessler (MP): design of an optimized method for smooth convex minimization, using SDPs.
- ’16 Lessard, Recht, Packard (SIOPT): smaller SDPs for linear convergence, via integral quadratic constraints (“IQCs”). Essentially Lyapunov functions.

In this presentation:

- ’17 T, Hendrickx and Glineur (MP): tightness and primal/dual interpretations of the certificates. (essentially previous slides)
- ’17 T, Hendrickx and Glineur (SIOPT): tightness of generalizations (see later).

PEP genealogy ("my humble, biased, view on...")

Base methodological developments:

'14 Drori and Teboulle (MP): upper bounds on worst-case behaviors of FO methods via SDP. Problems scale with number of iterations ($N \times N$ SDP matrices).

'16 Kim and Fessler (MP): design of an optimized method for smooth convex minimization, using SDPs.

'16 Lessard, Recht, Packard (SIOPT): smaller SDPs for linear convergence, via integral quadratic constraints ("IQCs"). Essentially Lyapunov functions.

In this presentation:

'17 T, Hendrickx and Glineur (MP): tightness and primal/dual interpretations of the certificates. (essentially previous slides)

'17 T, Hendrickx and Glineur (SIOPT): tightness of generalizations (see later).

— Other examples randomly picked from different works.

PEP genealogy ("my humble, biased, view on...")

Base methodological developments:

'14 Drori and Teboulle (MP): upper bounds on worst-case behaviors of FO methods via SDP. Problems scale with number of iterations ($N{\times}N$ SDP matrices).

'16 Kim and Fessler (MP): design of an optimized method for smooth convex minimization, using SDPs.

'16 Lessard, Recht, Packard (SIOPT): smaller SDPs for linear convergence, via integral quadratic constraints ("IQCs"). Essentially Lyapunov functions.

In this presentation:

'17 T, Hendrickx and Glineur (MP): tightness and primal/dual interpretations of the certificates. (essentially previous slides)

'17 T, Hendrickx and Glineur (SIOPT): tightness of generalizations (see later).

— Other examples randomly picked from different works.

'19 T, Bach (COLT): potential functions with tightness for sublinear convergence rates. Essentially: try to "force" simpler proofs. (if time allows)

PEP genealogy ("my humble, biased, view on...")

Base methodological developments:

'14 Drori and Teboulle (MP): upper bounds on worst-case behaviors of FO methods via SDP. Problems scale with number of iterations ($N \times N$ SDP matrices).

'16 Kim and Fessler (MP): design of an optimized method for smooth convex minimization, using SDPs.

'16 Lessard, Recht, Packard (SIOPT): smaller SDPs for linear convergence, via integral quadratic constraints ("IQCs"). Essentially Lyapunov functions.

In this presentation:

'17 T, Hendrickx and Glineur (MP): tightness and primal/dual interpretations of the certificates. (essentially previous slides)

'17 T, Hendrickx and Glineur (SIOPT): tightness of generalizations (see later).

— Other examples randomly picked from different works.

'19 T, Bach (COLT): potential functions with tightness for sublinear convergence rates. Essentially: try to "force" simpler proofs. (if time allows)

But also:

PEP genealogy (“my humble, biased, view on...”)

Base methodological developments:

'14 Drori and Teboulle (MP): upper bounds on worst-case behaviors of FO methods via SDP. Problems scale with number of iterations ($N\times N$ SDP matrices).

'16 Kim and Fessler (MP): design of an optimized method for smooth convex minimization, using SDPs.

'16 Lessard, Recht, Packard (SIOPT): smaller SDPs for linear convergence, via integral quadratic constraints (“IQCs”). Essentially Lyapunov functions.

In this presentation:

'17 T, Hendrickx and Glineur (MP): tightness and primal/dual interpretations of the certificates. (essentially previous slides)

'17 T, Hendrickx and Glineur (SIOPT): tightness of generalizations (see later).

— Other examples randomly picked from different works.

'19 T, Bach (COLT): potential functions with tightness for sublinear convergence rates. Essentially: try to “force” simpler proofs. (if time allows)

But also:

◇ Fair amount of algorithmic analyses (and design) originated from SDPs (from different authors, examples below), in different settings.

PEP genealogy ("my humble, biased, view on...")

Base methodological developments:

- '14 Drori and Teboulle (MP): upper bounds on worst-case behaviors of FO methods via SDP. Problems scale with number of iterations ($N \times N$ SDP matrices).
- '16 Kim and Fessler (MP): design of an optimized method for smooth convex minimization, using SDPs.
- '16 Lessard, Recht, Packard (SIOPT): smaller SDPs for linear convergence, via integral quadratic constraints ("IQCs"). Essentially Lyapunov functions.

In this presentation:

- '17 T, Hendrickx and Glineur (MP): tightness and primal/dual interpretations of the certificates. (essentially previous slides)
- '17 T, Hendrickx and Glineur (SIOPT): tightness of generalizations (see later).
- — Other examples randomly picked from different works.
- '19 T, Bach (COLT): potential functions with tightness for sublinear convergence rates. Essentially: try to "force" simpler proofs. (if time allows)

But also:

- ◇ Fair amount of algorithmic analyses (and design) originated from SDPs (from different authors, examples below), in different settings.
- ◇ We try keeping track of related works in the toolbox' manual (see later).

Toy example

Performance estimation

Further examples

Toward simpler proofs

Conclusions and discussions

Performance estimation problems

The approach we used for the gradient method can be used for a variety of methods.

Performance estimation problems

The approach we used for the gradient method can be used for a variety of methods.

Some attractive features of the approach:

Performance estimation problems

The approach we used for the gradient method can be used for a variety of methods.

Some attractive features of the approach:

- any primal solution is a lower bound (i.e., a function),

Performance estimation problems

The approach we used for the gradient method can be used for a variety of methods.

Some attractive features of the approach:

- any primal solution is a lower bound (i.e., a function),
- any dual solution is a worst-case guarantee (i.e., a proof),

Performance estimation problems

The approach we used for the gradient method can be used for a variety of methods.

Some attractive features of the approach:

- any primal solution is a lower bound (i.e., a function),
- any dual solution is a worst-case guarantee (i.e., a proof),
- it can be solved using semidefinite programming (SDP).

Classes of problems

Constrained and regularized optimization problems can be handled, as well:

$$\min_{x \in \mathbb{R}^d} f(x) + h(x),$$

for different functional classes:

Classes of problems

Constrained and regularized optimization problems can be handled, as well:

$$\min_{x \in \mathbb{R}^d} f(x) + h(x),$$

for different functional classes:

- different types of (smooth or non-smooth) convex functions,

Classes of problems

Constrained and regularized optimization problems can be handled, as well:

$$\min_{x\in\mathbb{R}^d} f(x) + h(x),$$

for different functional classes:

- different types of (smooth or non-smooth) convex functions,
- convex indicator and support functions,

Classes of problems

Constrained and regularized optimization problems can be handled, as well:

$$\min_{x \in \mathbb{R}^d} f(x) + h(x),$$

for different functional classes:

- different types of (smooth or non-smooth) convex functions,
- convex indicator and support functions,
- non-convex smooth functions,

Classes of problems

Constrained and regularized optimization problems can be handled, as well:

$$\min_{x \in \mathbb{R}^d} f(x) + h(x),$$

for different functional classes:

- different types of (smooth or non-smooth) convex functions,
- convex indicator and support functions,
- non-convex smooth functions,
- problem classes whose interpolation conditions are SDP-representable:

Classes of problems

Constrained and regularized optimization problems can be handled, as well:

$$\min_{x \in \mathbb{R}^d} f(x) + h(x),$$

for different functional classes:

- different types of (smooth or non-smooth) convex functions,
- convex indicator and support functions,
- non-convex smooth functions,
- problem classes whose interpolation conditions are SDP-representable:
  e.g., monotone inclusions, variational inequalities, fixed-point problems.

Algorithms

The approach can be used to obtain (tight) results for variety of "fixed-step" methods:

Algorithms

The approach can be used to obtain (tight) results for variety of "fixed-step" methods:

- (sub)gradient methods,

Algorithms

The approach can be used to obtain (tight) results for variety of "fixed-step" methods:

- (sub)gradient methods,
- inexact gradients methods,

Algorithms

The approach can be used to obtain (tight) results for variety of “fixed-step” methods:

- (sub)gradient methods,
- inexact gradients methods,
- proximal point methods,
- projected and proximal gradients methods,
- mirror descent,

Algorithms

The approach can be used to obtain (tight) results for variety of "fixed-step" methods:

- (sub)gradient methods,
- inexact gradients methods,
- proximal point methods,
- projected and proximal gradients methods,
- mirror descent,
- conditional gradient methods,
- splitting methods,
- randomized/stochastic gradient methods,
- distributed/decentralized gradient methods.

Algorithms

The approach can be used to obtain (tight) results for variety of “fixed-step” methods:

- (sub)gradient methods,
- inexact gradients methods,
- proximal point methods,
- projected and proximal gradients methods,
- mirror descent,
- conditional gradient methods,
- splitting methods,
- randomized/stochastic gradient methods,
- distributed/decentralized gradient methods.

Those includes fast/accelerated variants.

Algorithms

The approach can be used to obtain (tight) results for variety of "fixed-step" methods:

- (sub)gradient methods,
- inexact gradients methods,
- proximal point methods,
- projected and proximal gradients methods,
- mirror descent,
- conditional gradient methods,
- splitting methods,
- randomized/stochastic gradient methods,
- distributed/decentralized gradient methods.

Those includes fast/accelerated variants.

The approach usually provides upper bounds (no a priori tightness) in other situations.

Algorithms

The approach can be used to obtain (tight) results for variety of "fixed-step" methods:

- (sub)gradient methods,
- inexact gradients methods,
- proximal point methods,
- projected and proximal gradients methods,
- mirror descent,
- conditional gradient methods,
- splitting methods,
- randomized/stochastic gradient methods,
- distributed/decentralized gradient methods.

Those includes fast/accelerated variants.

The approach usually provides upper bounds (no a priori tightness) in other situations.

SDPs might scale badly, for example in stochastic or distributed settings.

Convergence measures

Different convergence measures can be taken into account.

Among others:

Convergence measures

Different convergence measures can be taken into account.

Among others:

- $f(x_N) - f(x_\star)$, $\|x_N - x_\star\|^2$, $\|f'(x_N)\|^2$,

Convergence measures

Different convergence measures can be taken into account.

Among others:

- $f(x_N) - f(x_\star)$, $\|x_N - x_\star\|^2$, $\|f'(x_N)\|^2$,
- best iterates on the way:

$$\min_{0\leq i\leq N} f(x_i) - f(x_\star), \quad \min_{0\leq i\leq N} \|x_i - x_\star\|^2, \quad \min_{0\leq i\leq N} \left\|f'(x_i)\right\|^2,$$

Convergence measures

Different convergence measures can be taken into account.

Among others:

- $f(x_N) - f(x_\star)$, $\|x_N - x_\star\|^2$, $\|f'(x_N)\|^2$,
- best iterates on the way:

$$\min_{0\leq i\leq N} f(x_i) - f(x_\star), \quad \min_{0\leq i\leq N} \|x_i - x_\star\|^2, \quad \min_{0\leq i\leq N} \left\|f'(x_i)\right\|^2,$$

- any concave function of f_i's, $\left\langle x_i, g_j \right\rangle$'s, $\|g_i\|^2$'s and $\|x_i\|^2$'s.

Toy example

Performance estimation

Further examples

Toward simpler proofs

Conclusions and discussions

Gradient method: final words?

Gradient method: final words?

Question: Let $x_{k+1} = x_k - \frac{1}{L} f'(x_k)$; what is the smallest τ such that

$$f(x_N) - f_* \leq \tau \|x_0 - x_*\|^2$$

is valid, for all x_0 and all L-smooth and convex function f?

Gradient method: final words?

Question: Let $x_{k+1} = x_k - \frac{1}{L} f'(x_k)$; what is the smallest τ such that

$$f(x_N) - f_* \leq \tau \|x_0 - x_*\|^2$$

is valid, for all x_0 and all L-smooth and convex function f?

From (Drori and Teboulle, 2014):

$$\max \left\{ \frac{f(x_N) - f(x_\star)}{\|x_0 - x_*\|^2} \right\} = \frac{L}{4N+2}.$$

Gradient method: final words?

Question: Let $x_{k+1} = x_k - \frac{1}{L} f'(x_k)$; what is the smallest τ such that

$$f(x_N) - f_* \leq \tau \|x_0 - x_*\|^2$$

is valid, for all x_0 and all L-smooth and convex function f?

From (Drori and Teboulle, 2014):

$$\max \left\{ \frac{f(x_N) - f(x_\star)}{\|x_0 - x_*\|^2} \right\} = \frac{L}{4N+2}.$$

Observation: worst-cases achieved on one-dimensional Huber losses:

$$\min_{x \in \mathbb{R}} f(x) = \begin{cases} \frac{L}{2N+1} x - \frac{L}{2(2N+1)^2} & \text{when } \|x\| \geq \frac{1}{2N+1} \\ \frac{L}{2} x^2 & \text{otherwise,} \end{cases}$$

Numerically observed from trace norm minimization heuristic.

François Glineur
(UCLouvain)

Etienne de Klerk
(Tilburg & Delft)

"On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions" (2017, Opt. Letters)

Steepest descent with inexact search directions

$$\min_{x \in \mathbb{R}^d} f(x),$$

with $f \in \mathcal{F}_{\mu,L}$ (L-smooth μ-strongly convex).

Steepest descent with inexact search directions

$$\min_{x \in \mathbb{R}^d} f(x),$$

with $f \in \mathcal{F}_{\mu,L}$ (L-smooth μ-strongly convex).

Relative error model:

$$\|f'(\mathbf{x}_i) - \mathbf{d}_i\| \leq \varepsilon \|f'(\mathbf{x}_i)\| \quad i = 0, 1, \ldots, \tag{1}$$

Steepest descent with inexact search directions

$$\min_{x \in \mathbb{R}^d} f(x),$$

with $f \in \mathcal{F}_{\mu,L}$ (L-smooth μ-strongly convex).

Relative error model:

$$\|f'(\mathbf{x}_i) - \mathbf{d}_i\| \leq \varepsilon \|f'(\mathbf{x}_i)\| \quad i = 0, 1, \ldots, \tag{1}$$

Noisy gradient descent method with exact line search

Input: $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$, $\mathbf{x}_0 \in \mathbb{R}^n$, $0 \leq \varepsilon < 1$.

for $i = 0, 1, \ldots$

Select any seach direction $\mathbf{d}_i$ that satisfies (1);

$\gamma = \text{argmin}_{\gamma \in \mathbb{R}} f(\mathbf{x}_i - \gamma \mathbf{d}_i)$

$\mathbf{x}_{i+1} = \mathbf{x}_i - \gamma \mathbf{d}_i$

Steepest descent with inexact search directions

$$\min_{x \in \mathbb{R}^d} f(x),$$

with $f \in \mathcal{F}_{\mu,L}$ (L-smooth μ-strongly convex).

Relative error model:

$$\|f'(\mathbf{x}_i) - \mathbf{d}_i\| \leq \varepsilon \|f'(\mathbf{x}_i)\| \quad i = 0, 1, \ldots, \tag{1}$$

Noisy gradient descent method with exact line search

Input: $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$, $\mathbf{x}_0 \in \mathbb{R}^n$, $0 \leq \varepsilon < 1$.

for $i = 0, 1, \ldots$

Select any seach direction $\mathbf{d}_i$ that satisfies (1);

$\gamma = \mathrm{argmin}_{\gamma \in \mathbb{R}} f\left(\mathbf{x}_i - \gamma \mathbf{d}_i\right)$

$\mathbf{x}_{i+1} = \mathbf{x}_i - \gamma \mathbf{d}_i$

Worst-case behavior:

$$f(\mathbf{x}_{i+1}) - f_* \leq \left(\frac{1-\kappa_\varepsilon}{1+\kappa_\varepsilon}\right)^2 (f(\mathbf{x}_i) - f_*) \quad i = 0, 1, \ldots$$

where $\kappa_\varepsilon = \frac{\mu}{L}\frac{(1-\varepsilon)}{(1+\varepsilon)}$.

Steepest descent with inexact search directions

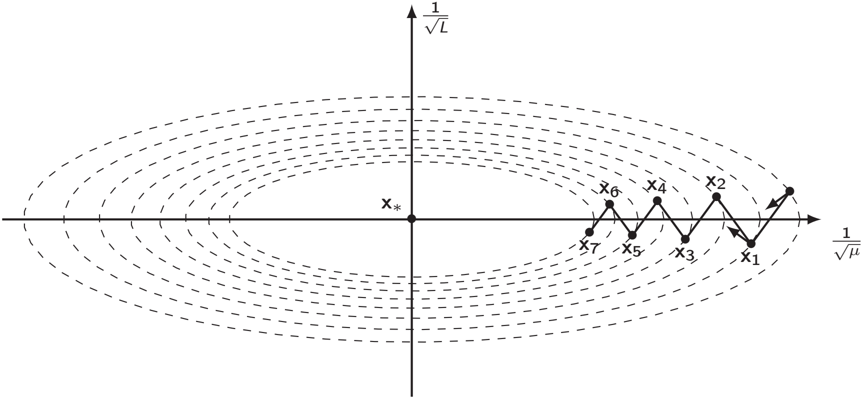
Quadratic worst-case function:

$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n}\lambda_i x_i^2 \quad \text{where} \quad 0 < \mu = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n = L.$$

Steepest descent with inexact search directions

Quadratic worst-case function:

$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i x_i^2 \quad \text{where} \quad 0 < \mu = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n = L.$$



What does the proof look like?

Aggregate constraints:

What does the proof look like?

Aggregate constraints:

$$f_0 \geq f_1 + \langle g_1, x_0 - x_1 \rangle + \frac{1}{2L}\|g_0 - g_1\|^2 + \frac{\mu}{2\left(1 - \frac{\mu}{L}\right)}\|x_0 - x_1 - (g_0 - g_1)/L\|^2$$

$$f_\star \geq f_0 + \langle g_0, x_\star - x_0 \rangle + \frac{1}{2L}\|g_\star - g_0\|^2 + \frac{\mu}{2\left(1 - \frac{\mu}{L}\right)}\|x_\star - x_0 - (g_\star - g_0)/L\|^2$$

$$f_\star \geq f_1 + \langle g_1, x_\star - x_1 \rangle + \frac{1}{2L}\|g_\star - g_1\|^2 + \frac{\mu}{2\left(1 - \frac{\mu}{L}\right)}\|x_\star - x_1 - (g_\star - g_1)/L\|^2$$

$$0 = \langle g_0, g_1 \rangle$$

$$0 = \langle g_1, x_1 - x_0 \rangle$$

What does the proof look like?

Aggregate constraints:

$$
\begin{aligned}
f_0 &\geq f_1 + \langle g_1, x_0 - x_1 \rangle + \frac{1}{2L}\|g_0 - g_1\|^2 + \frac{\mu}{2\left(1 - \frac{\mu}{L}\right)}\|x_0 - x_1 - (g_0 - g_1)/L\|^2 \\
f_\star &\geq f_0 + \langle g_0, x_\star - x_0 \rangle + \frac{1}{2L}\|g_\star - g_0\|^2 + \frac{\mu}{2\left(1 - \frac{\mu}{L}\right)}\|x_\star - x_0 - (g_\star - g_0)/L\|^2 \\
f_\star &\geq f_1 + \langle g_1, x_\star - x_1 \rangle + \frac{1}{2L}\|g_\star - g_1\|^2 + \frac{\mu}{2\left(1 - \frac{\mu}{L}\right)}\|x_\star - x_1 - (g_\star - g_1)/L\|^2 \\
0 &= \langle g_0, g_1 \rangle \\
0 &= \langle g_1, x_1 - x_0 \rangle
\end{aligned}
$$

with multipliers

What does the proof look like?

Aggregate constraints:

$$f_0 \geq f_1 + \langle g_1, x_0 - x_1 \rangle + \frac{1}{2L}\|g_0 - g_1\|^2 + \frac{\mu}{2\left(1 - \frac{\mu}{L}\right)}\|x_0 - x_1 - (g_0 - g_1)/L\|^2$$

$$f_\star \geq f_0 + \langle g_0, x_\star - x_0 \rangle + \frac{1}{2L}\|g_\star - g_0\|^2 + \frac{\mu}{2\left(1 - \frac{\mu}{L}\right)}\|x_\star - x_0 - (g_\star - g_0)/L\|^2$$

$$f_\star \geq f_1 + \langle g_1, x_\star - x_1 \rangle + \frac{1}{2L}\|g_\star - g_1\|^2 + \frac{\mu}{2\left(1 - \frac{\mu}{L}\right)}\|x_\star - x_1 - (g_\star - g_1)/L\|^2$$

$$0 = \langle g_0, g_1 \rangle$$

$$0 = \langle g_1, x_1 - x_0 \rangle$$

with multipliers

$$y_1 = \frac{L - \mu}{L + \mu}, \quad y_2 = 2\mu \frac{(L - \mu)}{(L + \mu)^2}, \quad y_3 = \frac{2\mu}{L + \mu}, \quad y_4 = \frac{2}{L + \mu}, \quad y_5 = 1.$$

What does the proof look like?

Resulting inequality:

$$
\begin{aligned}
f_1 - f_\star \quad \leq \quad & \left(\tfrac{L-\mu}{L+\mu}\right)^2 (f_0 - f_\star) \\
& - \tfrac{\mu L(L+3\mu)}{2(L+\mu)^2} \left\| x_0 - \tfrac{L+\mu}{L+3\mu} x_1 - \tfrac{2\mu}{L+3\mu} x_\star - \tfrac{3L+\mu}{L^2+3\mu L} g_0 - \tfrac{L+\mu}{L^2+3\mu L} g_1 \right\|^2 \\
& - \tfrac{2L\mu^2}{L^2+2L\mu-3\mu^2} \left\| x_1 - x_\star - \tfrac{(L-\mu)^2}{2\mu L(L+\mu)} g_0 - \tfrac{L+\mu}{2\mu L} g_1 \right\|^2 .
\end{aligned}
$$

What does the proof look like?

Resulting inequality:

$$\begin{aligned} f_1 - f_\star \quad \leq \quad & \left(\tfrac{L-\mu}{L+\mu}\right)^2 (f_0 - f_\star) \\ & - \tfrac{\mu L(L+3\mu)}{2(L+\mu)^2} \left\| x_0 - \tfrac{L+\mu}{L+3\mu} x_1 - \tfrac{2\mu}{L+3\mu} x_\star - \tfrac{3L+\mu}{L^2+3\mu L} g_0 - \tfrac{L+\mu}{L^2+3\mu L} g_1 \right\|^2 \\ & - \tfrac{2L\mu^2}{L^2+2L\mu-3\mu^2} \left\| x_1 - x_\star - \tfrac{(L-\mu)^2}{2\mu L(L+\mu)} g_0 - \tfrac{L+\mu}{2\mu L} g_1 \right\|^2 . \end{aligned}$$

Last two terms nonpositive, so

$$f_1 - f_\star \leq \left(\frac{L-\mu}{L+\mu}\right)^2 (f_0 - f_\star).$$

What does the proof look like?

Resulting inequality:

$$
\begin{aligned}
f_1 - f_\star \quad \leq \quad & \left(\tfrac{L-\mu}{L+\mu}\right)^2 (f_0 - f_\star) \\
& - \tfrac{\mu L(L+3\mu)}{2(L+\mu)^2} \left\| x_0 - \tfrac{L+\mu}{L+3\mu} x_1 - \tfrac{2\mu}{L+3\mu} x_\star - \tfrac{3L+\mu}{L^2+3\mu L} g_0 - \tfrac{L+\mu}{L^2+3\mu L} g_1 \right\|^2 \\
& - \tfrac{2L\mu^2}{L^2+2L\mu-3\mu^2} \left\| x_1 - x_\star - \tfrac{(L-\mu)^2}{2\mu L(L+\mu)} g_0 - \tfrac{L+\mu}{2\mu L} g_1 \right\|^2.
\end{aligned}
$$

Last two terms nonpositive, so

$$f_1 - f_\star \leq \left(\frac{L-\mu}{L+\mu}\right)^2 (f_0 - f_\star).$$

One actually has equality at optimality, due to the quadratic example.

Yoel Drori
(Google)

"Efficient first-order methods for convex minimization: a constructive approach" (2019, MP)

Optimized gradient methods

Smooth convex minimization setting:

$$\min_{x \in \mathbb{R}^d} f(x)$$

with f being L-smooth and convex, with black-box oracle $f'(.)$ available.

Optimized gradient methods

Smooth convex minimization setting:

$$\min_{x \in \mathbb{R}^d} f(x)$$

with f being L-smooth and convex, with black-box oracle $f'(.)$ available.

Lower bound for large-scale setting ($d \geq N+2$) by Drori (2017):

$$f(x_N) - f(x_\star) \geq \frac{L\|x_0 - x_\star\|^2}{2\theta_N^2},$$

with $\theta_0 = 1$, and:

$$\theta_{i+1} = \begin{cases} \frac{1+\sqrt{4\theta_i^2+1}}{2} & \text{if } i \leq N-2, \\ \frac{1+\sqrt{8\theta_i^2+1}}{2} & \text{if } i = N-1. \end{cases}$$

Optimized gradient methods

Smooth convex minimization setting:

$$\min_{x\in\mathbb{R}^d} f(x)$$

with f being L-smooth and convex, with black-box oracle $f'(.)$ available.

Lower bound for large-scale setting ($d \geq N+2$) by Drori (2017):

$$f(x_N) - f(x_\star) \geq \frac{L\|x_0 - x_\star\|^2}{2\theta_N^2} = O(1/N^2),$$

with $\theta_0 = 1$, and:

$$\theta_{i+1} = \begin{cases} \frac{1+\sqrt{4\theta_i^2+1}}{2} & \text{if } i \leq N-2, \\ \frac{1+\sqrt{8\theta_i^2+1}}{2} & \text{if } i = N-1. \end{cases}$$

Optimized gradient methods

Smooth convex minimization setting:

$$\min_{x\in\mathbb{R}^d} f(x)$$

with f being L-smooth and convex, with black-box oracle $f'(.)$ available.

Lower bound for large-scale setting ($d \geq N+2$) by Drori (2017):

$$f(x_N) - f(x_\star) \geq \frac{L\|x_0 - x_\star\|^2}{2\theta_N^2} = O(1/N^2),$$

with $\theta_0 = 1$, and:

$$\theta_{i+1} = \begin{cases} \frac{1+\sqrt{4\theta_i^2+1}}{2} & \text{if } i \leq N-2, \\ \frac{1+\sqrt{8\theta_i^2+1}}{2} & \text{if } i = N-1. \end{cases}$$

Coherent with historical lower bounds (Nemirovski & Yudin 1983) and optimal methods (Nemirovski 1982), (Nesterov 1983).

Optimized gradient methods

Three methods with the same (optimal) worst-case behavior

Greedy First-order Method (GFOM)

Inputs: f, x_0, N.

For $i = 1, 2, \ldots$

$$x_i = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(x) : \; x \in x_0 + \operatorname{span}\{f'(x_0), \ldots, f'(x_{i-1})\} \right\}.$$

Worst-case guarantee:

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

Optimized gradient methods

Three methods with the same (optimal) worst-case behavior

Optimized gradient method with exact line-search

Inputs: f, x_0, N.

For $i = 1, \ldots, N$

$$y_i = \left(1 - \frac{1}{\theta_i}\right) x_{i-1} + \frac{1}{\theta_i} x_0$$

$$d_i = \left(1 - \frac{1}{\theta_i}\right) f'(x_{i-1}) + \frac{1}{\theta_i} \left(2 \sum_{j=0}^{i-1} \theta_j f'(x_j)\right)$$

$$\alpha = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}}\ f(y_i + \alpha d_i)$$

$$x_i = y_i + \alpha d_i$$

Worst-case guarantee:

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

Optimized gradient methods

Three methods with the same (optimal) worst-case behavior

Optimized gradient method

Inputs: f, x_0, N.

For $i = 1, \ldots, N$

$$y_i = x_{i-1} - \frac{1}{L} f'(x_{i-1})$$

$$z_i = x_0 - \frac{2}{L} \sum_{j=0}^{i-1} \theta_j f'(x_j)$$

$$x_i = \left(1 - \frac{1}{\theta_i}\right) y_i + \frac{1}{\theta_i} z_i$$

Worst-case guarantee:

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

See also (Drori & Teboulle 2014) and (Kim & Fessler 2016).

What does the proof look like?

Aggregate quite a few constraints with appropriate coefficients.

What does the proof look like?

Aggregate quite a few constraints with appropriate coefficients.

Weighted sum can be rewritten exactly as (for the three cases):

$$f(x_N) - f(x_\star) \leq \frac{L\|x_0 - x_\star\|^2}{2\theta_N^2} - \frac{L}{2\theta_N^2}\left\| x_0 - x_* - \frac{\theta_N}{L} f'(x_N) - \frac{2}{L}\sum_{i=0}^{N-1} \theta_i f'(x_i) \right\|^2$$

Ernest Ryu
(UCLA)

Carolina Bergeling
(Lund)

Pontus Giselsson
(Lund)

"Operator splitting performance estimation: Tight contraction factors and optimal parameter selection" (2018, arXiv:1812.00146)

Douglas-Rachford Splitting I

Let f and h be two convex, closed, proper functions. (Overrelaxed) DRS for solving

$$\min_{x\in\mathbb{R}^d} f(x) + h(x),$$

consists in iterating:

Douglas-Rachford Splitting I

Let f and h be two convex, closed, proper functions. (Overrelaxed) DRS for solving

$$\min_{x \in \mathbb{R}^d} f(x) + h(x),$$

consists in iterating:

$$\begin{aligned}
x_{k+1} &= \operatorname{argmin}_{x \in \mathbb{R}^d} \{\gamma h(x) + \tfrac{1}{2}\|x - w_k\|^2\} \\
y_{k+1} &= \operatorname{argmin}_{y \in \mathbb{R}^d} \{\gamma f(y) + \tfrac{1}{2}\|y - 2x_{k+1} + w_k\|^2\} \\
w_{k+1} &= w_k + \theta(y_{k+1} - x_{k+1}),
\end{aligned}$$

for some choices of (θ, γ).

Douglas-Rachford Splitting II

Let A, and B be maximally monotone operators; and let $J_{\gamma A} := (I + \gamma A)^{-1}$ and $J_{\gamma B} := (I + \gamma B)^{-1}$ be their respective resolvents.

Douglas-Rachford Splitting II

Let A, and B be maximally monotone operators; and let $J_{\gamma A} := (I + \gamma A)^{-1}$ and $J_{\gamma B} := (I + \gamma B)^{-1}$ be their respective resolvents.

Monotone inclusion problem:

$$\underset{x \in \mathbb{R}^d}{\text{find}} \; 0 \in A(x) + B(x),$$

Douglas-Rachford Splitting II

Let A, and B be maximally monotone operators; and let $J_{\gamma A} := (I + \gamma A)^{-1}$ and $J_{\gamma B} := (I + \gamma B)^{-1}$ be their respective resolvents.

Monotone inclusion problem:

$$\operatorname*{find}_{x \in \mathbb{R}^d} \; 0 \in A(x) + B(x),$$

(overrelaxed) Douglas-Rachford for solving the monotone inclusion

$$w_{k+1} = (I - \theta J_{\gamma B} + \theta J_{\gamma A}(2J_{\gamma B} - I))w_k.$$

Douglas-Rachford Splitting II

Let A, and B be maximally monotone operators; and let $J_{\gamma A} := (I + \gamma A)^{-1}$ and $J_{\gamma B} := (I + \gamma B)^{-1}$ be their respective resolvents.

Monotone inclusion problem:

$$\underset{x \in \mathbb{R}^d}{\text{find}} \ 0 \in A(x) + B(x),$$

(overrelaxed) Douglas-Rachford for solving the monotone inclusion

$$w_{k+1} = (I - \theta J_{\gamma B} + \theta J_{\gamma A}(2J_{\gamma B} - I))w_k.$$

Recover optimization setting with $A = \partial f$ and $B = \partial h$.

Assumptions

Assumptions

Nontrivial rates by assuming something more on A and/or B.

Assumptions

Nontrivial rates by assuming something more on A and/or B.

Pick among the following (well documented) assumptions:

Assumptions

Nontrivial rates by assuming something more on A and/or B.

Pick among the following (well documented) assumptions:

- A convex function f is commonly assumed to be (for all $x, y \in \mathbb{R}^d$):
  - μ-strongly convex $\quad f(x) \geq f(y) + \langle \partial f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2,$
  - L-smooth $\quad f(x) \leq f(y) + \langle f'(y), x - y \rangle + \frac{L}{2}\|x - y\|^2.$

Assumptions

Nontrivial rates by assuming something more on A and/or B.

Pick among the following (well documented) assumptions:

- A convex function f is commonly assumed to be (for all $x, y \in \mathbb{R}^d$):
  - μ-strongly convex $\quad f(x) \geq f(y) + \langle \partial f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2$,
  - L-smooth $\quad f(x) \leq f(y) + \langle f'(y), x - y \rangle + \frac{L}{2}\|x - y\|^2$.

- A max. monotone operators B is commonly assumed to be (for all $x, y \in \mathbb{R}^d$):
  - a subdifferential $\quad B = \partial f(x)$,
  - μ-strongly monotone $\quad \langle B(x) - B(y), x - y \rangle \geq \mu \|x - y\|^2$,
  - β-cocoercive $\quad \langle B(x) - B(y), x - y \rangle \geq \beta \|B(x) - B(y)\|^2$,
  - L-Lipschitz $\quad \|B(x) - B(y)\| \leq L\|x - y\|$.

Contraction factor

Contraction factor

Question: When is the DRS iteration a contraction? What is the smallest ρ such that

$$\left\|w_1 - w_1'\right\| \le \rho \left\|w_0 - w_0'\right\|,$$

for all $w_0, w_0' \in \mathbb{R}^d$ and $w_1,\ w_1'$ generated with DRS from respectively w_0 and w_0'?

Contraction factor

Question: When is the DRS iteration a contraction? What is the smallest ρ such that

$$\|w_1 - w_1'\| \leq \rho \|w_0 - w_0'\|,$$

for all $w_0, w_0' \in \mathbb{R}^d$ and w_1, w_1' generated with DRS from respectively w_0 and w_0'?

Warning for the next few slides:

Contraction factor

Question: When is the DRS iteration a contraction? What is the smallest ρ such that

$$\|w_1 - w_1'\| \leq \rho \|w_0 - w_0'\|,$$

for all $w_0, w_0' \in \mathbb{R}^d$ and w_1, w_1' generated with DRS from respectively w_0 and w_0'?

Warning for the next few slides:

- the expressions are horrible,

Contraction factor

Question: When is the DRS iteration a contraction? What is the smallest ρ such that

$$\left\|w_1 - w_1'\right\| \leq \rho\left\|w_0 - w_0'\right\|,$$

for all $w_0, w_0' \in \mathbb{R}^d$ and $w_1,\, w_1'$ generated with DRS from respectively w_0 and w_0'?

Warning for the next few slides:

- ◇ the expressions are horrible,
- ◇ barely obtainable by hand,

Contraction factor

Question: When is the DRS iteration a contraction? What is the smallest ρ such that

$$\|w_1 - w_1'\| \leq \rho \|w_0 - w_0'\|,$$

for all $w_0, w_0' \in \mathbb{R}^d$ and w_1, w_1' generated with DRS from respectively w_0 and w_0'?

Warning for the next few slides:

- the expressions are horrible,
- barely obtainable by hand,
- computer-generated (Mathematica or Matlab),

Contraction factor

Question: When is the DRS iteration a contraction? What is the smallest ρ such that

$$\|w_1 - w_1'\| \leq \rho \|w_0 - w_0'\|,$$

for all $w_0, w_0' \in \mathbb{R}^d$ and w_1, w_1' generated with DRS from respectively w_0 and w_0'?

Warning for the next few slides:

- the expressions are horrible,
- barely obtainable by hand,
- computer-generated (Mathematica or Matlab),
- verifiable by hand (possibly long algebraic proofs).

Contraction factor

Question: When is the DRS iteration a contraction? What is the smallest ρ such that

$$\|w_1 - w_1'\| \le \rho \|w_0 - w_0'\|,$$

for all $w_0, w_0' \in \mathbb{R}^d$ and w_1, w_1' generated with DRS from respectively w_0 and w_0'?

Warning for the next few slides:

- the expressions are horrible,
- barely obtainable by hand,
- computer-generated (Mathematica or Matlab),
- verifiable by hand (possibly long algebraic proofs).

Intuitions can be developed, but this is another story ☺

DRS contraction factors

Table: Contraction factors for DRS: assumptions beyond max. monotonicity.

#	Properties for A	Properties for B	Reference	Sharp	Notes
O1	∂f, f: str. cvx & smooth	∂g	[1,2]	✔	
O2	∂f, f: str. cvx	∂g, g: smooth	[3]	✖	1.
M1	str. mono. & cocoercive	-	[3]	✔	
M2	str. mono. & Lipschitz	-	[3]	✔	2.
M3	str. mono.	cocoercive	[3]	✖	
M4	str. mono.	Lipschitz	[4]	✖	3.

1. sharp rates for some parameter choices in [3]
2. Lions and Mercier [5] provided conservative rate in this setting
3. sharp rate when B is skew linear in [4]

[1] Giselsson, Boyd, Diagonal Scaling in DRS and ADMM, 2014.
[2] Giselsson, Boyd, Linear Convergence and Metric Selection in DRS and ADMM, 2017.
[3] Giselsson, Tight Global Linear Convergence Rate Bounds for DRS, 2017.
[4] Moursi, Vandenberghe. DRS for a Lipschitz continuous and a strongly monotone operator, 2018.
[5] Lions, Mercier. Splitting Algorithms for the Sum of Two Nonlinear Operators, 1979.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

We have $\|Tx - Ty\| \leq \rho\|x - y\|$ for all $x, y \in \mathcal{H}$ with:

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

We have $\|Tx - Ty\| \leq \rho\|x - y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \left\{ \right.$$

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

We have $\|Tx - Ty\| \leq \rho \|x - y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} \left|1 - \theta \frac{\beta}{\beta+1}\right| & \text{if } \mu\beta - \mu + \beta < 0, \text{ and } \theta \leq 2\frac{(\beta+1)(\mu-\beta-\mu\beta)}{\mu+\mu\beta-\beta-\beta^2-2\mu\beta^2}, \\ & \\ & \end{cases}$$

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

We have $\|Tx - Ty\| \leq \rho\|x-y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} \left|1 - \theta\frac{\beta}{\beta+1}\right| & \text{if } \mu\beta - \mu + \beta < 0, \text{ and } \theta \leq 2\frac{(\beta+1)(\mu-\beta-\mu\beta)}{\mu+\mu\beta-\beta-\beta^2-2\mu\beta^2}, \\ \left|1 - \theta\frac{1+\mu\beta}{(\mu+1)(\beta+1)}\right| & \text{if } \mu\beta - \mu - \beta > 0, \text{ and } \theta \leq 2\frac{\mu^2+\beta^2+\mu\beta+\mu+\beta-\mu^2\beta^2}{\mu^2+\beta^2+\mu^2\beta+\mu\beta^2+\mu+\beta-2\mu^2\beta^2}, \\ & \\ & \end{cases}$$

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

We have $\|Tx - Ty\| \leq \rho\|x - y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} |1 - \theta \frac{\beta}{\beta+1}| & \text{if } \mu\beta - \mu + \beta < 0, \text{ and } \theta \leq 2\frac{(\beta+1)(\mu-\beta-\mu\beta)}{\mu+\mu\beta-\beta-\beta^2-2\mu\beta^2}, \\ |1 - \theta \frac{1+\mu\beta}{(\mu+1)(\beta+1)}| & \text{if } \mu\beta - \mu - \beta > 0, \text{ and } \theta \leq 2\frac{\mu^2+\beta^2+\mu\beta+\mu+\beta-\mu^2\beta^2}{\mu^2+\beta^2+\mu^2\beta+\mu\beta^2+\mu+\beta-2\mu^2\beta^2}, \\ |1 - \theta| & \text{if } \theta \geq 2\frac{\mu\beta+\mu+\beta}{2\mu\beta+\mu+\beta}, \\ & \end{cases}$$

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

We have $\|Tx - Ty\| \leq \rho\|x - y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} |1 - \theta\frac{\beta}{\beta+1}| & \text{if } \mu\beta - \mu + \beta < 0, \text{ and } \theta \leq 2\frac{(\beta+1)(\mu-\beta-\mu\beta)}{\mu+\mu\beta-\beta-\beta^2-2\mu\beta^2}, \\ |1 - \theta\frac{1+\mu\beta}{(\mu+1)(\beta+1)}| & \text{if } \mu\beta - \mu - \beta > 0, \text{ and } \theta \leq 2\frac{\mu^2+\beta^2+\mu\beta+\mu+\beta-\mu^2\beta^2}{\mu^2+\beta^2+\mu^2\beta+\mu\beta^2+\mu+\beta-2\mu^2\beta^2}, \\ |1 - \theta| & \text{if } \theta \geq 2\frac{\mu\beta+\mu+\beta}{2\mu\beta+\mu+\beta}, \\ |1 - \theta\frac{\mu}{\mu+1}| & \text{if } \mu\beta + \mu - \beta < 0, \text{ and } \theta \leq 2\frac{(\mu+1)(\beta-\mu-\mu\beta)}{\beta+\mu\beta-\mu-\mu^2-2\mu^2\beta}, \end{cases}$$

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

We have $\|Tx - Ty\| \leq \rho\|x - y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} |1 - \theta\frac{\beta}{\beta+1}| & \text{if } \mu\beta - \mu + \beta < 0, \text{ and } \theta \leq 2\frac{(\beta+1)(\mu-\beta-\mu\beta)}{\mu+\mu\beta-\beta-\beta^2-2\mu\beta^2}, \\ |1 - \theta\frac{1+\mu\beta}{(\mu+1)(\beta+1)}| & \text{if } \mu\beta - \mu - \beta > 0, \text{ and } \theta \leq 2\frac{\mu^2+\beta^2+\mu\beta+\mu+\beta-\mu^2\beta^2}{\mu^2+\beta^2+\mu^2\beta+\mu\beta^2+\mu+\beta-2\mu^2\beta^2}, \\ |1 - \theta| & \text{if } \theta \geq 2\frac{\mu\beta+\mu+\beta}{2\mu\beta+\mu+\beta}, \\ |1 - \theta\frac{\mu}{\mu+1}| & \text{if } \mu\beta + \mu - \beta < 0, \text{ and } \theta \leq 2\frac{(\mu+1)(\beta-\mu-\mu\beta)}{\beta+\mu\beta-\mu-\mu^2-2\mu^2\beta}, \\ X & \text{otherwise,} \end{cases}$$

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

We have $\|Tx - Ty\| \le \rho\|x-y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} |1 - \theta\frac{\beta}{\beta+1}| & \text{if } \mu\beta - \mu + \beta < 0, \text{ and } \theta \le 2\frac{(\beta+1)(\mu-\beta-\mu\beta)}{\mu+\mu\beta-\beta-\beta^2-2\mu\beta^2}, \\ |1 - \theta\frac{1+\mu\beta}{(\mu+1)(\beta+1)}| & \text{if } \mu\beta - \mu - \beta > 0, \text{ and } \theta \le 2\frac{\mu^2+\beta^2+\mu\beta+\mu+\beta-\mu^2\beta^2}{\mu^2+\beta^2+\mu^2\beta+\mu\beta^2+\mu+\beta-2\mu^2\beta^2}, \\ |1-\theta| & \text{if } \theta \ge 2\frac{\mu\beta+\mu+\beta}{2\mu\beta+\mu+\beta}, \\ |1 - \theta\frac{\mu}{\mu+1}| & \text{if } \mu\beta + \mu - \beta < 0, \text{ and } \theta \le 2\frac{(\mu+1)(\beta-\mu-\mu\beta)}{\beta+\mu\beta-\mu-\mu^2-2\mu^2\beta}, \\ X & \text{otherwise,} \end{cases}$$

with

$$X = \frac{\sqrt{2-\theta}}{2}\sqrt{\frac{((2-\theta)\mu(\beta+1)-\theta\beta(\mu-1))\,((2-\theta)\beta(\mu+1)-\theta\mu(\beta-1))}{(2-\theta)\mu\beta(\mu+1)(\beta+1)-\theta\mu^2\beta^2}}.$$

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

We have $\|Tx - Ty\| \leq \rho\|x - y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} |1 - \theta\frac{\beta}{\beta+1}| & \text{if } \mu\beta - \mu + \beta < 0, \text{ and } \theta \leq 2\frac{(\beta+1)(\mu-\beta-\mu\beta)}{\mu+\mu\beta-\beta-\beta^2-2\mu\beta^2}, \\ |1 - \theta\frac{1+\mu\beta}{(\mu+1)(\beta+1)}| & \text{if } \mu\beta - \mu - \beta > 0, \text{ and } \theta \leq 2\frac{\mu^2+\beta^2+\mu\beta+\mu+\beta-\mu^2\beta^2}{\mu^2+\beta^2+\mu^2\beta+\mu\beta^2+\mu+\beta-2\mu^2\beta^2}, \\ |1 - \theta| & \text{if } \theta \geq 2\frac{\mu\beta+\mu+\beta}{2\mu\beta+\mu+\beta}, \\ |1 - \theta\frac{\mu}{\mu+1}| & \text{if } \mu\beta + \mu - \beta < 0, \text{ and } \theta \leq 2\frac{(\mu+1)(\beta-\mu-\mu\beta)}{\beta+\mu\beta-\mu-\mu^2-2\mu^2\beta}, \\ X & \text{otherwise,} \end{cases}$$

with

$$X = \frac{\sqrt{2-\theta}}{2}\sqrt{\frac{((2-\theta)\mu(\beta+1)-\theta\beta(\mu-1))\,((2-\theta)\beta(\mu+1)-\theta\mu(\beta-1))}{(2-\theta)\mu\beta(\mu+1)(\beta+1)-\theta\mu^2\beta^2}}.$$

- ◇ The first four cases are achieved on 1-dimensional examples (primal is simpler).

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

We have $\|Tx - Ty\| \leq \rho \|x - y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} |1 - \theta \frac{\beta}{\beta+1}| & \text{if } \mu\beta - \mu + \beta < 0, \text{ and } \theta \leq 2\frac{(\beta+1)(\mu-\beta-\mu\beta)}{\mu+\mu\beta-\beta-\beta^2-2\mu\beta^2}, \\ |1 - \theta \frac{1+\mu\beta}{(\mu+1)(\beta+1)}| & \text{if } \mu\beta - \mu - \beta > 0, \text{ and } \theta \leq 2\frac{\mu^2+\beta^2+\mu\beta+\mu+\beta-\mu^2\beta^2}{\mu^2+\beta^2+\mu^2\beta+\mu\beta^2+\mu+\beta-2\mu^2\beta^2}, \\ |1 - \theta| & \text{if } \theta \geq 2\frac{\mu\beta+\mu+\beta}{2\mu\beta+\mu+\beta}, \\ |1 - \theta \frac{\mu}{\mu+1}| & \text{if } \mu\beta + \mu - \beta < 0, \text{ and } \theta \leq 2\frac{(\mu+1)(\beta-\mu-\mu\beta)}{\beta+\mu\beta-\mu-\mu^2-2\mu^2\beta}, \\ X & \text{otherwise,} \end{cases}$$

with

$$X = \frac{\sqrt{2-\theta}}{2}\sqrt{\frac{((2-\theta)\mu(\beta+1)-\theta\beta(\mu-1))\,((2-\theta)\beta(\mu+1)-\theta\mu(\beta-1))}{(2-\theta)\mu\beta(\mu+1)(\beta+1)-\theta\mu^2\beta^2}}.$$

- The first four cases are achieved on 1-dimensional examples (primal is simpler).
- Fifth case is achieved on 2-dimensional example (dual is simpler).

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

Examples on which those bounds are attained?

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

Examples on which those bounds are attained?

- Case 1: (1-dimensional) $A = N_{\{0\}}$ (i.e., $J_{\lambda A} = 0$), $B = \frac{1}{\beta} I$ for $\rho = |1 - \theta \frac{\beta}{\beta+1}|$.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

Examples on which those bounds are attained?

- Case 1: (1-dimensional) $A = N_{\{0\}}$ (i.e., $J_{\lambda A} = 0$), $B = \frac{1}{\beta} I$ for $\rho = |1 - \theta \frac{\beta}{\beta+1}|$.
- Case 2: (1-dimensional) $A = \mu I$, $B = \frac{1}{\beta} I$ for $\rho = |1 - \theta \frac{1+\mu\beta}{(\mu+1)(\beta+1)}|$.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

Examples on which those bounds are attained?

- Case 1: (1-dimensional) $A = N_{\{0\}}$ (i.e., $J_{\lambda A} = 0$), $B = \frac{1}{\beta} I$ for $\rho = |1 - \theta \frac{\beta}{\beta+1}|$.
- Case 2: (1-dimensional) $A = \mu I$, $B = \frac{1}{\beta} I$ for $\rho = |1 - \theta \frac{1+\mu\beta}{(\mu+1)(\beta+1)}|$.
- Case 3: (1-dimensional) $A = N_{\{0\}}$, $B = 0$ for $\rho = |1 - \theta|$.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

Examples on which those bounds are attained?

- Case 1: (1-dimensional) $A = N_{\{0\}}$ (i.e., $J_{\lambda A} = 0$), $B = \frac{1}{\beta} I$ for $\rho = |1 - \theta \frac{\beta}{\beta+1}|$.
- Case 2: (1-dimensional) $A = \mu I$, $B = \frac{1}{\beta} I$ for $\rho = |1 - \theta \frac{1+\mu\beta}{(\mu+1)(\beta+1)}|$.
- Case 3: (1-dimensional) $A = N_{\{0\}}$, $B = 0$ for $\rho = |1 - \theta|$.
- Case 4: (1-dimensional) $A = \mu I$, $B = 0$ for $\rho = |1 - \theta \frac{\mu}{\mu+1}|$.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B β-cocoercive.

Examples on which those bounds are attained?

- Case 1: (1-dimensional) $A = N_{\{0\}}$ (i.e., $J_{\lambda A} = 0$), $B = \frac{1}{\beta} I$ for $\rho = |1 - \theta \frac{\beta}{\beta+1}|$.
- Case 2: (1-dimensional) $A = \mu I$, $B = \frac{1}{\beta} I$ for $\rho = |1 - \theta \frac{1+\mu\beta}{(\mu+1)(\beta+1)}|$.
- Case 3: (1-dimensional) $A = N_{\{0\}}$, $B = 0$ for $\rho = |1 - \theta|$.
- Case 4: (1-dimensional) $A = \mu I$, $B = 0$ for $\rho = |1 - \theta \frac{\mu}{\mu+1}|$.
- Case 5: (2-dimensional) for appropriate (complicated) values of a and K:

$$A = \begin{pmatrix} \mu & -a \\ a & \mu \end{pmatrix}, \qquad B = \begin{pmatrix} \beta K & -\sqrt{K - K^2\beta^2} \\ \sqrt{K - K^2\beta^2} & \beta K \end{pmatrix},$$

for $\rho = \frac{\sqrt{2-\theta}}{2} \sqrt{\frac{((2-\theta)\mu(\beta+1) - \theta\beta(\mu-1))\,((2-\theta)\beta(\mu+1) - \theta\mu(\beta-1))}{(2-\theta)\mu\beta(\mu+1)(\beta+1) - \theta\mu^2\beta^2}}$.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B L-Lipschitz and monotone.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B L-Lipschitz and monotone.

We have $\|Tx - Ty\| \leq \rho\|x - y\|$ for all $x, y \in \mathcal{H}$ with:

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B L-Lipschitz and monotone.

We have $\|Tx - Ty\| \leq \rho \|x - y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} \dfrac{\theta + \sqrt{\frac{(2(\theta-1)\mu+\theta-2)^2 + L^2(\theta-2(\mu+1))^2}{L^2+1}}}{2(\mu+1)} & \text{if } (a), \\ \left|1 - \theta \frac{L+\mu}{(\mu+1)(L+1)}\right| & \text{if } (b), \\ \sqrt{\frac{(2-\theta)}{4\mu(L^2+1)} \frac{\left(\theta(L^2+1) - 2\mu(\theta+L^2-1)\right)\left(\theta\left(1+2\mu+L^2\right) - 2(\mu+1)\left(L^2+1\right)\right)}{2\mu(\theta+L^2-1) - (2-\theta)(1-L^2)}} & \text{otherwise,} \end{cases}$$

with

(a) $\mu \frac{-(2(\theta-1)\mu+\theta-2) + L^2(\theta-2(1+\mu))}{\sqrt{(2(\theta-1)\mu+\theta-2)^2 + L^2(\theta-2(\mu+1))^2}} \leq \sqrt{L^2+1}$,

(b) $L < 1$, $\mu > \frac{L^2+1}{(L-1)^2}$, and $\theta \leq \frac{2(\mu+1)(L+1)(\mu+\mu L^2 - L^2 - 2\mu L - 1)}{2\mu^2 - \mu + \mu L^3 - L^3 - 3\mu L^2 - L^2 - 2\mu^2 L - \mu L - L - 1}$.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B L-Lipschitz and monotone.

We have $\|Tx - Ty\| \leq \rho\|x - y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} \dfrac{\theta + \sqrt{\frac{(2(\theta-1)\mu+\theta-2)^2 + L^2(\theta-2(\mu+1))^2}{L^2+1}}}{2(\mu+1)} & \text{if } (a), \\ \left|1 - \theta\,\frac{L+\mu}{(\mu+1)(L+1)}\right| & \text{if } (b), \\ \sqrt{\frac{(2-\theta)}{4\mu(L^2+1)}\,\frac{\left(\theta(L^2+1)-2\mu(\theta+L^2-1)\right)\left(\theta\left(1+2\mu+L^2\right)-2(\mu+1)\left(L^2+1\right)\right)}{2\mu(\theta+L^2-1)-(2-\theta)(1-L^2)}} & \text{otherwise,} \end{cases}$$

with

(a) $\mu\,\frac{-(2(\theta-1)\mu+\theta-2)+L^2(\theta-2(1+\mu))}{\sqrt{(2(\theta-1)\mu+\theta-2)^2+L^2(\theta-2(\mu+1))^2}} \leq \sqrt{L^2+1}$,

(b) $L < 1$, $\mu > \frac{L^2+1}{(L-1)^2}$, and $\theta \leq \frac{2(\mu+1)(L+1)(\mu+\mu L^2-L^2-2\mu L-1)}{2\mu^2-\mu+\mu L^3-L^3-3\mu L^2-L^2-2\mu^2 L-\mu L-L-1}$.

◇ First and third cases are achieved on 2-dimensional examples (dual is simpler),

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B L-Lipschitz and monotone.

We have $\|Tx - Ty\| \leq \rho\|x-y\|$ for all $x, y \in \mathcal{H}$ with:

$$\rho = \begin{cases} \dfrac{\theta+\sqrt{\frac{(2(\theta-1)\mu+\theta-2)^2+L^2(\theta-2(\mu+1))^2}{L^2+1}}}{2(\mu+1)} & \text{if } (a), \\ \left|1-\theta\,\frac{L+\mu}{(\mu+1)(L+1)}\right| & \text{if } (b), \\ \sqrt{\frac{(2-\theta)}{4\mu(L^2+1)}\,\frac{\left(\theta(L^2+1)-2\mu(\theta+L^2-1)\right)\left(\theta\left(1+2\mu+L^2\right)-2(\mu+1)\left(L^2+1\right)\right)}{2\mu(\theta+L^2-1)-(2-\theta)(1-L^2)}} & \text{otherwise,} \end{cases}$$

with

(a) $\mu\frac{-(2(\theta-1)\mu+\theta-2)+L^2(\theta-2(1+\mu))}{\sqrt{(2(\theta-1)\mu+\theta-2)^2+L^2(\theta-2(\mu+1))^2}} \leq \sqrt{L^2+1}$,

(b) $L<1$, $\mu>\frac{L^2+1}{(L-1)^2}$, and $\theta \leq \frac{2(\mu+1)(L+1)(\mu+\mu L^2-L^2-2\mu L-1)}{2\mu^2-\mu+\mu L^3-L^3-3\mu L^2-L^2-2\mu^2 L-\mu L-L-1}$.

- ◇ First and third cases are achieved on 2-dimensional examples (dual is simpler),
- ◇ Second case is achieved on 1-dimensional example (primal is simpler).

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B L-Lipschitz and monotone.

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B L-Lipschitz and monotone.

Examples on which those bounds are attained?

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B L-Lipschitz and monotone.

Examples on which those bounds are attained?

- Case 1: (2-dimensional) We choose (see also Moursi & Vandenberghe 2018)

$$A = \mu I + N_{\{0\}\times\mathbb{R}}, \quad B = L \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

for $\rho = \frac{\theta + \sqrt{\frac{(2(\theta-1)\mu+\theta-2)^2 + L^2(\theta-2(\mu+1))^2}{L^2+1}}}{2(\mu+1)}$

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B L-Lipschitz and monotone.

Examples on which those bounds are attained?

- ◇ Case 1: (2-dimensional) We choose (see also Moursi & Vandenberghe 2018)

$$A = \mu\, I + N_{\{0\}\times\mathbb{R}}, \quad B = L \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

for $\rho = \frac{\theta + \sqrt{\frac{(2(\theta-1)\mu+\theta-2)^2 + L^2(\theta-2(\mu+1))^2}{L^2+1}}}{2(\mu+1)}$

- ◇ Case 2: (1-dimensional) $A = \mu I$, $B = LI$ for $\rho = |1 - \theta \frac{L+\mu}{(\mu+1)(L+1)}|$

Douglas-Rachford Splitting

Assumptions: A μ-strongly monotone, B L-Lipschitz and monotone.

Examples on which those bounds are attained?

- Case 1: (2-dimensional) We choose (see also Moursi & Vandenberghe 2018)

$$A = \mu I + N_{\{0\}\times\mathbb{R}}, \quad B = L\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

for $\rho = \frac{\theta + \sqrt{\frac{(2(\theta-1)\mu+\theta-2)^2+L^2(\theta-2(\mu+1))^2}{L^2+1}}}{2(\mu+1)}$

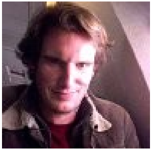
- Case 2: (1-dimensional) $A = \mu I$, $B = LI$ for $\rho = |1 - \theta\frac{L+\mu}{(\mu+1)(L+1)}|$
- Case 3: (2-dimensional) For appropriately chosen (complicated) K:

$$A = \mu I + N_{\mathbb{R}\times\{0\}}, \quad B = L\begin{pmatrix} K & -\sqrt{1-K^2} \\ \sqrt{1-K^2} & K \end{pmatrix},$$

for $\rho = \sqrt{\frac{(2-\theta)}{4\mu(L^2+1)}\frac{\big(\theta(L^2+1)-2\mu(\theta+L^2-1)\big)\big(\theta\big(1+2\mu+L^2\big)-2(\mu+1)\big(L^2+1\big)\big)}{2\mu(\theta+L^2-1)-(2-\theta)(1-L^2)}}.$

A-R. Dragomir
(ENS/TSE)

Jérôme Bolte
(TSE)



A. d'Aspremont
(CNRS/ENS)

"Optimal complexity and certification of Bregman first-order methods" (2019, arXiv:1911.08510)

Mirror descent/Bregman gradient/NoLips

Recall gradient descent with step size γ:

$$x_{k+1} = \operatorname*{argmin}_x \left\{ f(x_k) + \left\langle f'(x_k), x - x_k \right\rangle + \frac{1}{2\gamma} \|x - x_k\|^2 \right\}.$$

Mirror descent/Bregman gradient/NoLips

Recall gradient descent with step size γ:

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \left\{ f(x_k) + \left\langle f'(x_k), x - x_k \right\rangle + \frac{1}{2\gamma} \|x - x_k\|^2 \right\}.$$

High-level intuition: gradient descent should work well when

$$f(x_k) + \left\langle f'(x_k), x - x_k \right\rangle + \frac{1}{2\gamma} \|x - x_k\|^2$$

is a good approximation of f.

Mirror descent/Bregman gradient/NoLips

Recall gradient descent with step size γ:

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \left\{ f(x_k) + \left\langle f'(x_k), x - x_k \right\rangle + \frac{1}{2\gamma}\|x - x_k\|^2 \right\}.$$

High-level intuition: gradient descent should work well when

$$f(x_k) + \left\langle f'(x_k), x - x_k \right\rangle + \frac{1}{2\gamma}\|x - x_k\|^2$$

is a good approximation of f.

Mirror descent: change notion of distance and iterate:

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \left\{ f(x_k) + \left\langle f'(x_k), x - x_k \right\rangle + \frac{1}{\gamma} D_h(x, x_k) \right\}$$

where $D_h(x, x_k)$ is a Bregman divergence:

$$h(x) - h(x_k) - \left\langle h'(x_k), x - x_k \right\rangle \geq 0,$$

and h is strictly convex and differentiable.

Mirror descent/Bregman gradient/NoLips

Recent assumption for mirror descent: “relative smoothness” (Bauschke, Bolte, Teboulle, 2016), (Lu, Freund, Nesterov 2018):

$$Lh - f \text{ convex, } f \text{ convex, and } h \text{ strictly convex and differentiable}$$

(boils down to regular smoothness when $h = \frac{1}{2}\|.\|^2$).

Mirror descent/Bregman gradient/NoLips

Recent assumption for mirror descent: “relative smoothness” (Bauschke, Bolte, Teboulle, 2016), (Lu, Freund, Nesterov 2018):

$$Lh - f \text{ convex, } f \text{ convex, and } h \text{ strictly convex and differentiable}$$

(boils down to regular smoothness when $h = \frac{1}{2}\|.\|^2$).

Question: Let $x_{k+1} = \mathrm{MD}(x_k)$; what is the smallest τ such that

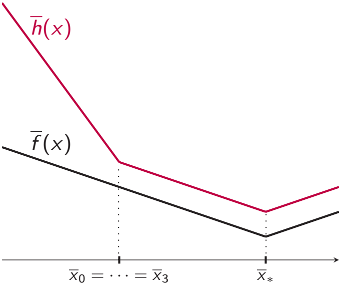
$$f(x_k) - f_* \leq \tau D_h(x_*, x_0)$$

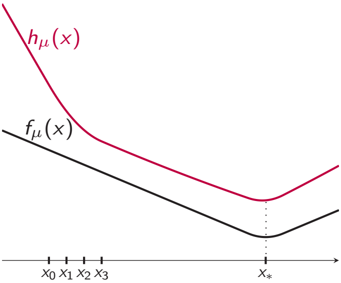
is valid, for all x_0, all (f, h) satisfying previous assumptions?

Mirror descent/Bregman gradient/NoLips

In this case: strictly convex differentiable functions (i.e., open set of functions).

Pathological nonsmooth limiting behaviors in the closure of this open set (via PEPs):





The guarantee

$$f(x_k) - f_* \leq \frac{LD_h(x_*, x_0)}{k}$$

cannot be improved (attained on example above).

Mirror descent/Bregman gradient/NoLips

Mirror descent/Bregman gradient/NoLips

Convexity of f, between x_* and x_i ($i = 0, \ldots, k$) with weight $\gamma_{*,i} = \frac{1}{k}$:

$$f(x_*) \geq f(x_i) + \left\langle f'(x_i), x_* - x_i \right\rangle,$$

Mirror descent/Bregman gradient/NoLips

Convexity of f, between x_* and x_i ($i = 0, \ldots, k$) with weight $\gamma_{*,i} = \frac{1}{k}$:

$$f(x_*) \geq f(x_i) + \left\langle f'(x_i), x_* - x_i \right\rangle,$$

convexity of f, between x_i and x_{i+1} ($i = 0, \ldots, k-1$) with weight $\gamma_{i,i+1} = \frac{i}{k}$:

$$f(x_i) \geq f(x_{i+1}) + \left\langle f'(x_{i+1}), x_i - x_{i+1} \right\rangle,$$

Mirror descent/Bregman gradient/NoLips

Convexity of f, between x_* and x_i ($i = 0, \ldots, k$) with weight $\gamma_{*,i} = \frac{1}{k}$:

$$f(x_*) \geq f(x_i) + \left\langle f'(x_i), x_* - x_i \right\rangle,$$

convexity of f, between x_i and x_{i+1} ($i = 0, \ldots, k-1$) with weight $\gamma_{i,i+1} = \frac{i}{k}$:

$$f(x_i) \geq f(x_{i+1}) + \left\langle f'(x_{i+1}), x_i - x_{i+1} \right\rangle,$$

convexity of $Lh - f$, between x_* and x_k with weight $\mu_{*,k} = \frac{1}{k}$:

$$Lh(x_*) - f(x_*) \geq Lh(x_k) - f(x_k) + \left\langle Lh'(x_k) - f'(x_k), x_* - x_k \right\rangle,$$

Mirror descent/Bregman gradient/NoLips

Convexity of f, between x_* and x_i $(i = 0, \ldots, k)$ with weight $\gamma_{*,i} = \frac{1}{k}$:

$$f(x_*) \geq f(x_i) + \left\langle f'(x_i), x_* - x_i \right\rangle,$$

convexity of f, between x_i and x_{i+1} $(i = 0, \ldots, k-1)$ with weight $\gamma_{i,i+1} = \frac{i}{k}$:

$$f(x_i) \geq f(x_{i+1}) + \left\langle f'(x_{i+1}), x_i - x_{i+1} \right\rangle,$$

convexity of $Lh - f$, between x_* and x_k with weight $\mu_{*,k} = \frac{1}{k}$:

$$Lh(x_*) - f(x_*) \geq Lh(x_k) - f(x_k) + \left\langle Lh'(x_k) - f'(x_k), x_* - x_k \right\rangle,$$

convexity of $Lh - f$, between x_{i+1} and x_i $(i = 0, \ldots, k-1)$ with weight $\mu_{i+1,i} = \frac{i+1}{k}$

$$Lh(x_{i+1}) - f(x_{i+1}) \geq Lh(x_i) - f(x_i) + \left\langle Lh'(x_i) - f'(x_i), x_{i+1} - x_i \right\rangle,$$

Mirror descent/Bregman gradient/NoLips

Convexity of f, between x_* and x_i ($i = 0, \ldots, k$) with weight $\gamma_{*,i} = \frac{1}{k}$:

$$f(x_*) \geq f(x_i) + \left\langle f'(x_i), x_* - x_i \right\rangle,$$

convexity of f, between x_i and x_{i+1} ($i = 0, \ldots, k-1$) with weight $\gamma_{i,i+1} = \frac{i}{k}$:

$$f(x_i) \geq f(x_{i+1}) + \left\langle f'(x_{i+1}), x_i - x_{i+1} \right\rangle,$$

convexity of $Lh - f$, between x_* and x_k with weight $\mu_{*,k} = \frac{1}{k}$:

$$Lh(x_*) - f(x_*) \geq Lh(x_k) - f(x_k) + \left\langle Lh'(x_k) - f'(x_k), x_* - x_k \right\rangle,$$

convexity of $Lh - f$, between x_{i+1} and x_i ($i = 0, \ldots, k-1$) with weight $\mu_{i+1,i} = \frac{i+1}{k}$

$$Lh(x_{i+1}) - f(x_{i+1}) \geq Lh(x_i) - f(x_i) + \left\langle Lh'(x_i) - f'(x_i), x_{i+1} - x_i \right\rangle,$$

convexity of $Lh - f$, between x_i and x_{i+1} ($i = 0, \ldots, k-1$) with weight $\mu_{i,i+1} = \frac{i}{k}$

$$Lh(x_i) - f(x_i) \geq Lh(x_{i+1}) - f(x_{i+1}) + \left\langle Lh'(x_{i+1}) - f'(x_{i+1}), x_i - x_{i+1} \right\rangle.$$

Mirror descent/Bregman gradient/NoLips

Convexity of f, between x_* and x_i ($i = 0, \ldots, k$) with weight $\gamma_{*,i} = \frac{1}{k}$:

$$f(x_*) \geq f(x_i) + \left\langle f'(x_i), x_* - x_i \right\rangle,$$

convexity of f, between x_i and x_{i+1} ($i = 0, \ldots, k-1$) with weight $\gamma_{i,i+1} = \frac{i}{k}$:

$$f(x_i) \geq f(x_{i+1}) + \left\langle f'(x_{i+1}), x_i - x_{i+1} \right\rangle,$$

convexity of $Lh - f$, between x_* and x_k with weight $\mu_{*,k} = \frac{1}{k}$:

$$Lh(x_*) - f(x_*) \geq Lh(x_k) - f(x_k) + \left\langle Lh'(x_k) - f'(x_k), x_* - x_k \right\rangle,$$

convexity of $Lh - f$, between x_{i+1} and x_i ($i = 0, \ldots, k-1$) with weight $\mu_{i+1,i} = \frac{i+1}{k}$

$$Lh(x_{i+1}) - f(x_{i+1}) \geq Lh(x_i) - f(x_i) + \left\langle Lh'(x_i) - f'(x_i), x_{i+1} - x_i \right\rangle,$$

convexity of $Lh - f$, between x_i and x_{i+1} ($i = 0, \ldots, k-1$) with weight $\mu_{i,i+1} = \frac{i}{k}$

$$Lh(x_i) - f(x_i) \geq Lh(x_{i+1}) - f(x_{i+1}) + \left\langle Lh'(x_{i+1}) - f'(x_{i+1}), x_i - x_{i+1} \right\rangle.$$

and reformulate:

$$f(x_k) - f(x_*) \leq L \frac{h(x_*) - h(x_0) - \left\langle h'(x_0), x_* - x_0 \right\rangle}{k},$$

where there is no residual term to neglect!

Avoiding semidefinite programming modeling steps?

Avoiding semidefinite programming modeling steps?

François Glineur
(UCLouvain)

Julien Hendrickx
(UCLouvain)

"Performance Estimation Toolbox (PESTO): automated worst-case analysis of first-order optimization methods" (CDC 2017)

PESTO example: contraction factors for DRS

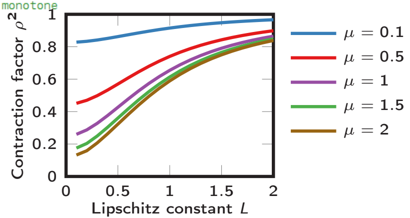
```
% (0) Initialize an empty PEP
P=pep();

N = 1;
% (1) Set up the class of monotone inclusions
paramA.L  =  1; paramA.mu = 0; % A is 1-Lipschitz and 0-strongly monotone
paramB.mu = .1;                % B is .1-strongly monotone

A = P.DeclareFunction('LipschitzStronglyMonotone',paramA);
B = P.DeclareFunction('StronglyMonotone',paramB);

w  = cell(N+1,1);   wp = cell(N+1,1);
x  = cell(N,1);     xp = cell(N,1);
y  = cell(N,1);     yp = cell(N,1);

% (2) Set up the starting points
w{1}    = P.StartingPoint(); wp{1}   = P.StartingPoint();
P.InitialCondition((w{1}-wp{1})^2<=1);

% (3) Algorithm
lambda = 1.3;       % step size (in the resolvents)
theta  = .9;         % overrelaxation

for k = 1 : N
    x{k}    = proximal_step(w{k},B,lambda);
    y{k}    = proximal_step(2*x{k}-w{k},A,lambda);
    w{k+1}  = w{k}-theta*(x{k}-y{k});

    xp{k}   = proximal_step(wp{k},B,lambda);
    yp{k}   = proximal_step(2*xp{k}-wp{k},A,lambda);
    wp{k+1}   = wp{k}-theta*(xp{k}-yp{k});
end

% (4) Set up the performance measure: ||z0-z1||^2
P.PerformanceMetric((w{k+1}-wp{k+1})^2);

% (5) Solve the PEP
P.solve()

% (6) Evaluate the output
double((w{k+1}-wp{k+1})^2)   % worst-case contraction factor
```

PESTO example: contraction factors for DRS

```
% (0) Initialize an empty PEP
P=pep();

N = 1;
% (1) Set up the class of monotone inclusions
paramA.L  =  1; paramA.mu = 0; % A is 1-Lipschitz and 0-strongly monotone
paramB.mu = .1;                % B is .1-strongly monotone

A = P.DeclareFunction('LipschitzStronglyMonotone',paramA);
B = P.DeclareFunction('StronglyMonotone',paramB);

w  = cell(N+1,1);   wp = cell(N+1,1);
x  = cell(N,1);     xp = cell(N,1);
y  = cell(N,1);     yp = cell(N,1);

% (2) Set up the starting points
w{1}    = P.StartingPoint(); wp{1}   = P.StartingPoint();
P.InitialCondition((w{1}-wp{1})^2<=1);

% (3) Algorithm
lambda = 1.3;       % step size (in the resolvents)
theta  = .9;         % overrelaxation

x{k}     = proximal_step(w{k},B,lambda);
y{k}     = proximal_step(2*x{k}-w{k},A,lambda);
w{k+1}   = w{k}-theta*(x{k}-y{k});
    xp{k}    = proximal_step(wp{k},B,lambda);
    yp{k}    = proximal_step(2*xp{k}-wp{k},A,lambda);
    wp{k+1}    = wp{k}-theta*(xp{k}-yp{k});
end

% (4) Set up the performance measure: ||z0-z1||^2
P.PerformanceMetric((w{k+1}-wp{k+1})^2);

% (5) Solve the PEP
P.solve()

% (6) Evaluate the output
double((w{k+1}-wp{k+1})^2)   % worst-case contraction factor
```

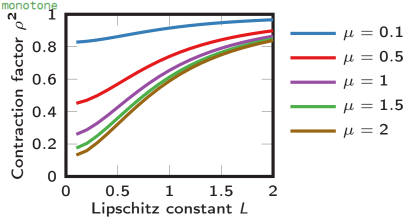
PESTO example: contraction factors for DRS

```
% (0) Initialize an empty PEP
P=pep();

N = 1;
% (1) Set up the class of monotone inclusions
paramA.L  =  1; paramA.mu = 0; % A is 1-Lipschitz and 0-strongly monotone
paramB.mu = .1;                % B is .1-strongly monotone

A = P.DeclareFunction('LipschitzStronglyMonotone',paramA);
B = P.DeclareFunction('StronglyMonotone',paramB);

w  = cell(N+1,1);   wp = cell(N+1,1);
x  = cell(N,1);     xp = cell(N,1);
y  = cell(N,1);     yp = cell(N,1);

% (2) Set up the starting points
w{1}    = P.StartingPoint(); wp{1}    = P.StartingPoint();
P.InitialCondition((w{1}-wp{1})^2<=1);

% (3) Algorithm
lambda = 1.3;        % step size (in the resolvents)
theta  = .9;          % overrelaxation

x{k}     = proximal_step(w{k},B,lambda);
y{k}     = proximal_step(2*x{k}-w{k},A,lambda);
w{k+1}   = w{k}-theta*(x{k}-y{k});
    xp{k}    = proximal_step(wp{k},B,lambda);
    yp{k}    = proximal_step(2*xp{k}-wp{k},A,lambda);
    wp{k+1}    = wp{k}-theta*(xp{k}-yp{k});
end

% (4) Set up the performance measure: ||z0-z1||^2
P.PerformanceMetric((w{k+1}-wp{k+1})^2);

% (5) Solve the PEP
P.solve()

% (6) Evaluate the output
double((w{k+1}-wp{k+1})^2)   % worst-case contraction factor
```



PESTO example: contraction factors for DRS

```
% (0) Initialize an empty PEP
P=pep();

N = 1;
% (1) Set up the class of monotone inclusions
paramA.L  =  1; paramA.mu = 0; % A is 1-Lipschitz and 0-strongly monotone
paramB.mu = .1;                % B is .1-strongly monotone

A = P.DeclareFunction('LipschitzStronglyMonotone',paramA);
B = P.DeclareFunction('StronglyMonotone',paramB);

w  = cell(N+1,1);   wp = cell(N+1,1);
x  = cell(N,1);     xp = cell(N,1);
y  = cell(N,1);     yp = cell(N,1);

% (2) Set up the starting points
w{1}    = P.StartingPoint(); wp{1}    = P.StartingPoint();
P.InitialCondition((w{1}-wp{1})^2<=1);

% (3) Algorithm
lambda = 1.3;       % step size (in the resolvents)
theta  = .9;         % overrelaxation

x{k}     = proximal_step(w{k},B,lambda);
y{k}     = proximal_step(2*x{k}-w{k},A,lambda);
w{k+1}   = w{k}-theta*(x{k}-y{k});
    xp{k}    = proximal_step(wp{k},B,lambda);
    yp{k}    = proximal_step(2*xp{k}-wp{k},A,lambda);
    wp{k+1}   = wp{k}-theta*(xp{k}-yp{k});
end

% (4) Set up the performance measure: ||z0-z1||^2
P.PerformanceMetric((w{k+1}-wp{k+1})^2);

% (5) Solve the PEP
P.solve()

% (6) Evaluate the output
double((w{k+1}-wp{k+1})^2)   % worst-case contraction factor
```



- ✔ fast prototyping ($\sim$ 20 effective lines)
- ✔ quick analyses ($\sim$ 10 minutes)
- ✔ computer-aided proofs (multipliers)

Current library of examples within PESTO

Includes... but not limited to

- subgradient, gradient, heavy-ball, fast gradient, optimized gradient methods,
- proximal point algorithm,
- projected and proximal gradient, accelerated/momentum versions,
- steepest descent, greedy/conjugate gradient methods,
- Douglas-Rachford/three operator splitting,
- Frank-Wolfe/conditional gradient,
- inexact gradient/fast gradient,
- Krasnoselskii-Mann and Halpern fixed-point iterations,
- mirror descent,
- stochastic methods: SAG, SAGA, SGD and variants.

Current library of examples within PESTO

Includes... but not limited to

- ◇ subgradient, gradient, heavy-ball, fast gradient, optimized gradient methods,
- ◇ proximal point algorithm,
- ◇ projected and proximal gradient, accelerated/momentum versions,
- ◇ steepest descent, greedy/conjugate gradient methods,
- ◇ Douglas-Rachford/three operator splitting,
- ◇ Frank-Wolfe/conditional gradient,
- ◇ inexact gradient/fast gradient,
- ◇ Krasnoselskii-Mann and Halpern fixed-point iterations,
- ◇ mirror descent,
- ◇ stochastic methods: SAG, SAGA, SGD and variants.

Current library of examples within PESTO

Includes... but not limited to

- subgradient, gradient, heavy-ball, fast gradient, optimized gradient methods,
- proximal point algorithm,
- projected and proximal gradient, accelerated/momentum versions,
- steepest descent, greedy/conjugate gradient methods,
- Douglas-Rachford/three operator splitting,
- Frank-Wolfe/conditional gradient,
- inexact gradient/fast gradient,
- Krasnoselskii-Mann and Halpern fixed-point iterations,
- mirror descent,
- stochastic methods: SAG, SAGA, SGD and variants.

PESTO contains most of the recent PEP-related advances (including techniques by other groups). Clean updated references in user manual.

Current library of examples within PESTO

Includes... but not limited to

- ◇ subgradient, gradient, heavy-ball, fast gradient, optimized gradient methods,
- ◇ proximal point algorithm,
- ◇ projected and proximal gradient, accelerated/momentum versions,
- ◇ steepest descent, greedy/conjugate gradient methods,
- ◇ Douglas-Rachford/three operator splitting,
- ◇ Frank-Wolfe/conditional gradient,
- ◇ inexact gradient/fast gradient,
- ◇ Krasnoselskii-Mann and Halpern fixed-point iterations,
- ◇ mirror descent,
- ◇ stochastic methods: SAG, SAGA, SGD and variants.

PESTO contains most of the recent PEP-related advances (including techniques by other groups). Clean updated references in user manual.

Among others, see works by Drori, Teboulle, Kim, Fessler, Ryu, Lieder, Lessard, Recht, Packard, Van Scoy, Cyrus, Gu, Yang, etc.

Toy example

Performance estimation

Further examples

Toward simpler proofs

Conclusions and discussions

Francis Bach
(Inria/ENS)

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

Some opinions on PEPs

Pros/cons of PEPs

Some opinions on PEPs

Pros/cons of PEPs

- ☺ Worst-case guarantees *cannot be improved*,

  details in (T, Hendrickx & Glineur 2017),

Some opinions on PEPs

Pros/cons of PEPs

- ☺ Worst-case guarantees *cannot be improved*,

  details in (T, Hendrickx & Glineur 2017),

- ☺ fair amount of generalizations (finite sums, constraints, prox, etc.),

  details in (T, Hendrickx & Glineur 2017); (Drori 2014), etc.

Some opinions on PEPs

Pros/cons of PEPs

- ☺ Worst-case guarantees *cannot be improved*,

  details in (T, Hendrickx & Glineur 2017),
- ☺ fair amount of generalizations (finite sums, constraints, prox, etc.),

  details in (T, Hendrickx & Glineur 2017); (Drori 2014), etc.
- ☹ SDPs typically become prohibitively large (with N and generalizations),

Some opinions on PEPs

Pros/cons of PEPs

- ☺ Worst-case guarantees *cannot be improved*,

  details in (T, Hendrickx & Glineur 2017),
- ☺ fair amount of generalizations (finite sums, constraints, prox, etc.),

  details in (T, Hendrickx & Glineur 2017); (Drori 2014), etc.
- ☹ SDPs typically become prohibitively large (with N and generalizations),
- ☹ proofs (may be) quite involved and hard to intuit,

Some opinions on PEPs

Pros/cons of PEPs

- ☺ Worst-case guarantees *cannot be improved*,

  details in (T, Hendrickx & Glineur 2017),

- ☺ fair amount of generalizations (finite sums, constraints, prox, etc.),

  details in (T, Hendrickx & Glineur 2017); (Drori 2014), etc.

- ☹ SDPs typically become prohibitively large (with N and generalizations),

- ☹ proofs (may be) quite involved and hard to intuit,

  examples in (Drori & Teboulle 2014), (Drori 2014), (Kim & Fessler 2016 2018), (Shi & Liu 2017), (de Klerk et al. 2017), etc.

Some opinions on PEPs

Pros/cons of PEPs

- ☺ Worst-case guarantees *cannot be improved*,

  details in (T, Hendrickx & Glineur 2017),
- ☺ fair amount of generalizations (finite sums, constraints, prox, etc.),

  details in (T, Hendrickx & Glineur 2017); (Drori 2014), etc.
- ☹ SDPs typically become prohibitively large (with N and generalizations),
- ☹ proofs (may be) quite involved and hard to intuit,

  examples in (Drori & Teboulle 2014), (Drori 2014), (Kim & Fessler 2016 2018), (Shi & Liu 2017), (de Klerk et al. 2017), etc.
- ☹ proofs (may be) hard to generalize (e.g., to handle projections, backtracking),

Some opinions on PEPs

Pros/cons of PEPs

- ☺ Worst-case guarantees *cannot be improved*,

  details in (T, Hendrickx & Glineur 2017),
- ☺ fair amount of generalizations (finite sums, constraints, prox, etc.),

  details in (T, Hendrickx & Glineur 2017); (Drori 2014), etc.
- ☹ SDPs typically become prohibitively large (with N and generalizations),
- ☹ proofs (may be) quite involved and hard to intuit,

  examples in (Drori & Teboulle 2014), (Drori 2014), (Kim & Fessler 2016 2018), (Shi & Liu 2017), (de Klerk et al. 2017), etc.
- ☹ proofs (may be) hard to generalize (e.g., to handle projections, backtracking),

  examples in (Kim & Fessler 2016 2018).

Some opinions on PEPs

Pros/cons of PEPs

- ☺ Worst-case guarantees *cannot be improved*,

  details in (T, Hendrickx & Glineur 2017),
- ☺ fair amount of generalizations (finite sums, constraints, prox, etc.),

  details in (T, Hendrickx & Glineur 2017); (Drori 2014), etc.
- ☹ SDPs typically become prohibitively large (with N and generalizations),
- ☹ proofs (may be) quite involved and hard to intuit,

  examples in (Drori & Teboulle 2014), (Drori 2014), (Kim & Fessler 2016 2018), (Shi & Liu 2017), (de Klerk et al. 2017), etc.
- ☹ proofs (may be) hard to generalize (e.g., to handle projections, backtracking),

  examples in (Kim & Fessler 2016 2018).
- ☺ allows reaching proofs that could barely be obtained by hand,

Some opinions on PEPs

Pros/cons of PEPs

- ☺ Worst-case guarantees *cannot be improved*,

  details in (T, Hendrickx & Glineur 2017),
- ☺ fair amount of generalizations (finite sums, constraints, prox, etc.),

  details in (T, Hendrickx & Glineur 2017); (Drori 2014), etc.
- ☹ SDPs typically become prohibitively large (with N and generalizations),
- ☹ proofs (may be) quite involved and hard to intuit,

  examples in (Drori & Teboulle 2014), (Drori 2014), (Kim & Fessler 2016 2018), (Shi & Liu 2017), (de Klerk et al. 2017), etc.
- ☹ proofs (may be) hard to generalize (e.g., to handle projections, backtracking),

  examples in (Kim & Fessler 2016 2018).
- ☺ allows reaching proofs that could barely be obtained by hand,
- ☺ easy to try via Performance EStimation TOolbox (PESTO),

Some opinions on PEPs

Pros/cons of PEPs

- ☺ Worst-case guarantees *cannot be improved*,

  details in (T, Hendrickx & Glineur 2017),
- ☺ fair amount of generalizations (finite sums, constraints, prox, etc.),

  details in (T, Hendrickx & Glineur 2017); (Drori 2014), etc.
- ☹ SDPs typically become prohibitively large (with N and generalizations),
- ☹ proofs (may be) quite involved and hard to intuit,

  examples in (Drori & Teboulle 2014), (Drori 2014), (Kim & Fessler 2016 2018), (Shi & Liu 2017), (de Klerk et al. 2017), etc.
- ☹ proofs (may be) hard to generalize (e.g., to handle projections, backtracking),

  examples in (Kim & Fessler 2016 2018).
- ☺ allows reaching proofs that could barely be obtained by hand,
- ☺ easy to try via Performance EStimation TOolbox (PESTO),
- ☺ possible to "force" simple proofs (typically at some cost: e.g., loosing tightness).

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 $\phi_{k+1}^f \leq \phi_k^f$ 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 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 $\phi^f_{k+1} \leq \phi^f_k$ 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:

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

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

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

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

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

Why is that nice? Very simple resulting proof:

$$N(f(x_N) - f_\star) \leq \phi_N^f \leq \phi_{N-1}^f \leq \ldots \leq \phi_0^f = \frac{L}{2}\|x_0 - x_\star\|^2,$$

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

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

How does it work for the gradient method?

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)

Starting point: candidate quadratic ϕ^f_k with *all the available information* at iteration k

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

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

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

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

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

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

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

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_{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 $\mathcal{V}_k$.)

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.

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^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 $\{a_k, b_k, c_k, d_k\}_k$.

In others words:

- *efficient (convex) representation of $\mathcal{V}_k$ available*!

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 $\{a_k, b_k, c_k, d_k\}_k$.

In others words:

- *efficient (convex) representation of $\mathcal{V}_k$ available*!
- idea: apply previous reformulation tricks to feasibility problem

$$0 \geq \max_f \phi^f_{k+1} - \phi^f_k.$$

The dual is also a feasibility problem, linear in $\{a_k, b_k, c_k, d_k\}_k$.

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

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

How does it work for the gradient method?

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

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

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

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

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

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

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?

How does it work for the gradient method?

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

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

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?

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

How does it work for the gradient method?

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

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

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:

How does it work for the gradient method?

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

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

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.

How does it work for the gradient method?

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

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

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

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

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

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

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:

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

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

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:

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

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

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:

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

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

How does it work for the gradient method?

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

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

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

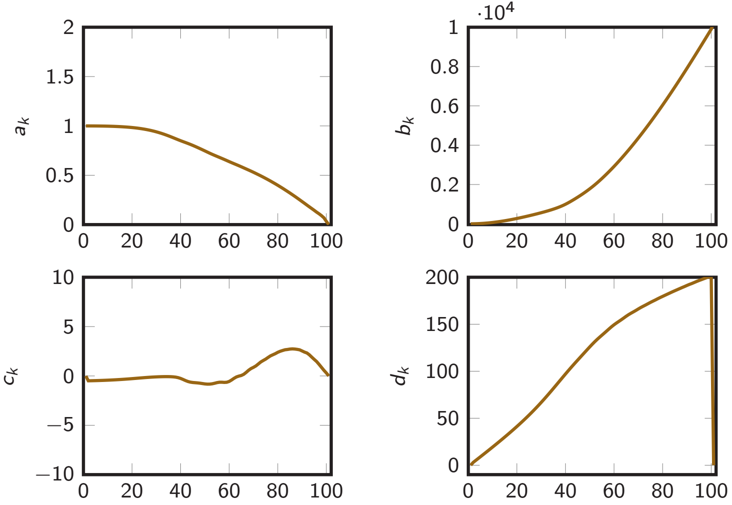
2. Observe the a_k, b_k, c_k, d_k's for some values of N.

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

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

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:

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

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

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

How does it work for the gradient method?

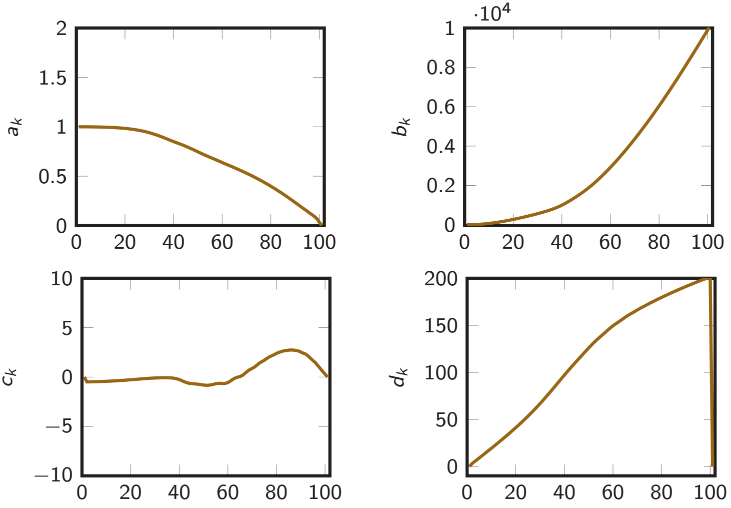
1. Solve the SDP for some values of N; recall final guarantee of the form:

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

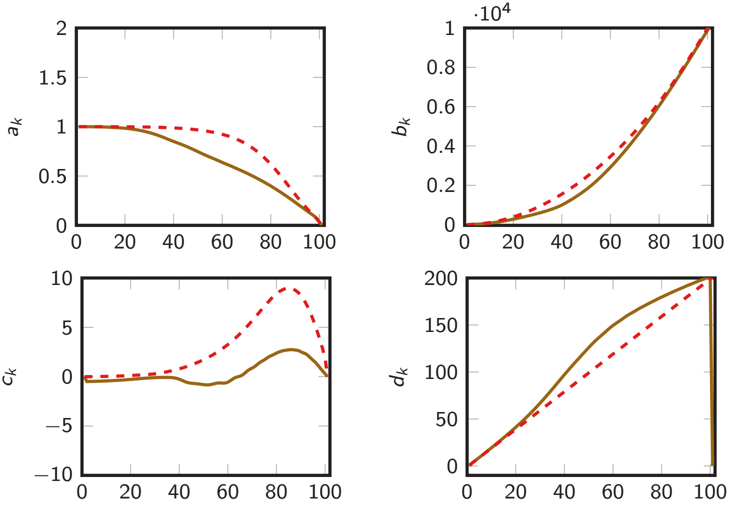
$$\begin{array}{rcccccc} N = & 1 & 2 & 3 & 4 & \ldots & 100 \\ b_N = & 4 & 9 & 16 & 25 & \ldots & 10201 \end{array}$$

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: $d_k = (2k+1)L$
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$
   Tentative simplification #3: $d_k = 0$

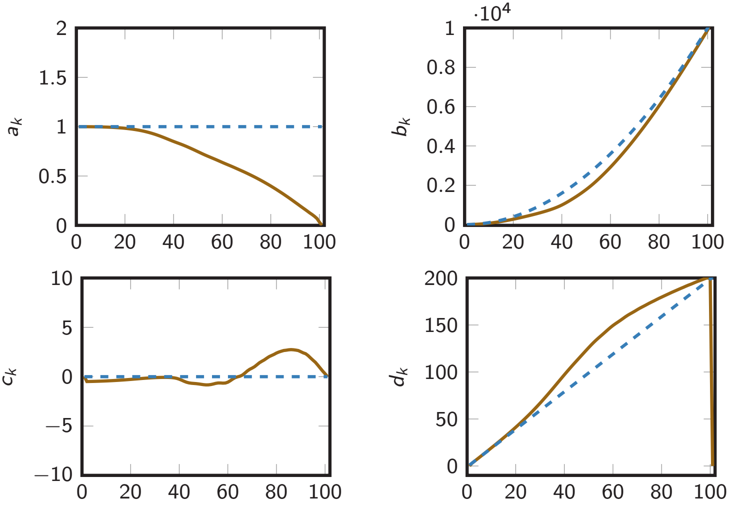
$$V_k = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top \left[\begin{pmatrix} a_k & c_k \\ c_k & b_k \end{pmatrix} \otimes I_d \right] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + d_k \left(f(x_k) - f(x_\star) \right)$$



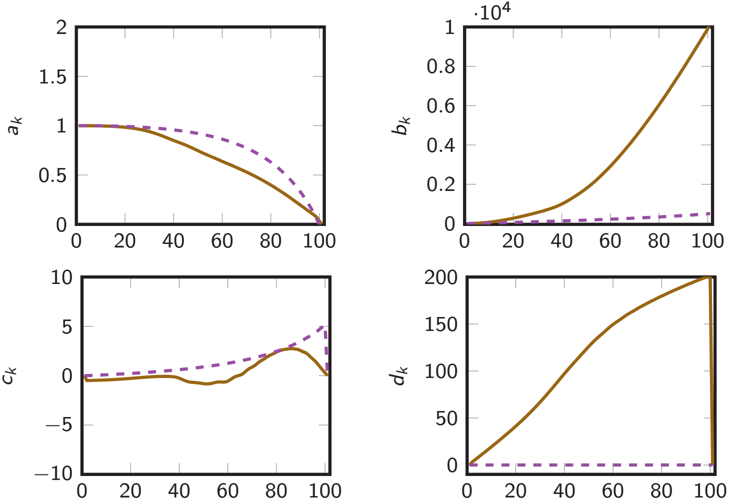
$$V_k = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top \left[\begin{pmatrix} a_k & c_k \\ c_k & b_k \end{pmatrix} \otimes I_d \right] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + (2k+1)L\left(f(x_k) - f(x_\star)\right)$$



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



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



How does it work for the gradient method?

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

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

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

   Tentative simplification #1: $d_k = (2k+1)L$
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$
   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:

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

$N =$	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.

   Tentative simplification #1: $d_k = (2k+1)L$ [success]
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$ [success]
   Tentative simplification #3: $d_k = 0$ [fail]

How does it work for the gradient method?

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

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

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

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: $d_k = (2k+1)L$ [success]
   Tentative simplification #2: $a_k = L^2$, $c_k = 0$ [success]
   Tentative simplification #3: $d_k = 0$ [fail]
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 $\|f'(x_N)\|^2 = O(N^{-2})$.

Potential functions

Simpler proof structures:

Potential functions

Simpler proof structures:

- ◇ allow keeping SDP formulations more tractable,

Potential functions

Simpler proof structures:

- allow keeping SDP formulations more tractable,
- hence usable with more complex settings (e.g., randomizations, stochasticity).

Potential functions

Simper proof structures:

- ◇ allow keeping SDP formulations more tractable,
- ◇ hence usable with more complex settings (e.g., randomizations, stochasticity).

More examples:

Potential functions

Simpler proof structures:

- ⋄ allow keeping SDP formulations more tractable,
- ⋄ hence usable with more complex settings (e.g., randomizations, stochasticity).

More examples:

- ⋄ all previous variants (everything that fits into regular PEPs)

Potential functions

Simpler proof structures:

- allow keeping SDP formulations more tractable,
- hence usable with more complex settings (e.g., randomizations, stochasticity).

More examples:

- all previous variants (everything that fits into regular PEPs)
- stochastic variants (e.g., finite sum, bounded variance, over-parametrization),

Potential functions

Simpler proof structures:

- ◇ allow keeping SDP formulations more tractable,
- ◇ hence usable with more complex settings (e.g., randomizations, stochasticity).

More examples:

- ◇ all previous variants (everything that fits into regular PEPs)
- ◇ stochastic variants (e.g., finite sum, bounded variance, over-parametrization),
- ◇ randomized block-coordinate variants,

Potential functions

Simpler proof structures:

- ◇ allow keeping SDP formulations more tractable,
- ◇ hence usable with more complex settings (e.g., randomizations, stochasticity).

More examples:

- ◇ all previous variants (everything that fits into regular PEPs)
- ◇ stochastic variants (e.g., finite sum, bounded variance, over-parametrization),
- ◇ randomized block-coordinate variants,

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

Toy example

Performance estimation

Further examples

Toward simpler proofs

Conclusions and discussions

Concluding remarks

Performance estimation's philosophy

Concluding remarks

Performance estimation's philosophy

- numerically allows obtaining tight bounds (rigorous baselines),

Concluding remarks

Performance estimation's philosophy

- numerically allows obtaining tight bounds (rigorous baselines),
- results can only be improved by changing algorithm and/or assumptions,

Concluding remarks

Performance estimation's philosophy

- numerically allows obtaining tight bounds (rigorous baselines),
- results can only be improved by changing algorithm and/or assumptions,
- helps designing analytical proofs (reduces to linear combinations of inequalities),

  proofs can be engineered using numerics & symbolic computations!

Concluding remarks

Performance estimation's philosophy

- ◇ numerically allows obtaining tight bounds (rigorous baselines),
- ◇ results can only be improved by changing algorithm and/or assumptions,
- ◇ helps designing analytical proofs (reduces to linear combinations of inequalities),
  proofs can be engineered using numerics & symbolic computations!
- ◇ fast prototyping:

Concluding remarks

Performance estimation's philosophy

- numerically allows obtaining tight bounds (rigorous baselines),
- results can only be improved by changing algorithm and/or assumptions,
- helps designing analytical proofs (reduces to linear combinations of inequalities),

  proofs can be engineered using numerics & symbolic computations!
- fast prototyping:

  before trying to prove your new FO method works; give PEP a try!

Concluding remarks

Performance estimation's philosophy

- numerically allows obtaining tight bounds (rigorous baselines),
- results can only be improved by changing algorithm and/or assumptions,
- helps designing analytical proofs (reduces to linear combinations of inequalities),

  proofs can be engineered using numerics & symbolic computations!
- fast prototyping:

  before trying to prove your new FO method works; give PEP a try!
- step forward to "reproducible theory" (useful for reviewing, too ☺).

Concluding remarks

Performance estimation's philosophy

- numerically allows obtaining tight bounds (rigorous baselines),
- results can only be improved by changing algorithm and/or assumptions,
- helps designing analytical proofs (reduces to linear combinations of inequalities),

  proofs can be engineered using numerics & symbolic computations!
- fast prototyping:

  before trying to prove your new FO method works; give PEP a try!
- step forward to "reproducible theory" (useful for reviewing, too ☺).

Difficulties:

Concluding remarks

Performance estimation's philosophy

- numerically allows obtaining tight bounds (rigorous baselines),
- results can only be improved by changing algorithm and/or assumptions,
- helps designing analytical proofs (reduces to linear combinations of inequalities),
  proofs can be engineered using numerics & symbolic computations!
- fast prototyping:
  before trying to prove your new FO method works; give PEP a try!
- step forward to "reproducible theory" (useful for reviewing, too ☺).

Difficulties:

- suffers from standard caveats of worst-case analyses,
  key is to find good assumptions/parametrization
- closed-form solutions might be involved.

Concluding remarks

Performance estimation's philosophy

- ◇ numerically allows obtaining tight bounds (rigorous baselines),
- ◇ results can only be improved by changing algorithm and/or assumptions,
- ◇ helps designing analytical proofs (reduces to linear combinations of inequalities),

  proofs can be engineered using numerics & symbolic computations!
- ◇ fast prototyping:

  before trying to prove your new FO method works; give PEP a try!
- ◇ step forward to "reproducible theory" (useful for reviewing, too ☺).

Difficulties:

- ◇ suffers from standard caveats of worst-case analyses,

  key is to find good assumptions/parametrization
- ◇ closed-form solutions might be involved.

Ongoing research directions, open questions:

Concluding remarks

Performance estimation's philosophy

- numerically allows obtaining tight bounds (rigorous baselines),
- results can only be improved by changing algorithm and/or assumptions,
- helps designing analytical proofs (reduces to linear combinations of inequalities),
  proofs can be engineered using numerics & symbolic computations!
- fast prototyping:
  before trying to prove your new FO method works; give PEP a try!
- step forward to "reproducible theory" (useful for reviewing, too ☺).

Difficulties:

- suffers from standard caveats of worst-case analyses,
  key is to find good assumptions/parametrization
- closed-form solutions might be involved.

Ongoing research directions, open questions:

- computer-assisted algorithmic design,

Concluding remarks

Performance estimation's philosophy

- ◇ numerically allows obtaining tight bounds (rigorous baselines),
- ◇ results can only be improved by changing algorithm and/or assumptions,
- ◇ helps designing analytical proofs (reduces to linear combinations of inequalities),
  proofs can be engineered using numerics & symbolic computations!
- ◇ fast prototyping:
  before trying to prove your new FO method works; give PEP a try!
- ◇ step forward to "reproducible theory" (useful for reviewing, too ☺).

Difficulties:

- ◇ suffers from standard caveats of worst-case analyses,
  key is to find good assumptions/parametrization
- ◇ closed-form solutions might be involved.

Ongoing research directions, open questions:

- ◇ computer-assisted algorithmic design,
- ◇ adaptive & structure-exploiting methods,

Concluding remarks

Performance estimation's philosophy

- ◇ numerically allows obtaining tight bounds (rigorous baselines),
- ◇ results can only be improved by changing algorithm and/or assumptions,
- ◇ helps designing analytical proofs (reduces to linear combinations of inequalities),
  proofs can be engineered using numerics & symbolic computations!
- ◇ fast prototyping:
  before trying to prove your new FO method works; give PEP a try!
- ◇ step forward to "reproducible theory" (useful for reviewing, too ☺).

Difficulties:

- ◇ suffers from standard caveats of worst-case analyses,
  key is to find good assumptions/parametrization
- ◇ closed-form solutions might be involved.

Ongoing research directions, open questions:

- ◇ computer-assisted algorithmic design,
- ◇ adaptive & structure-exploiting methods,
- ◇ non-convex & non-Euclidean settings?

Concluding remarks

Performance estimation's philosophy

- numerically allows obtaining tight bounds (rigorous baselines),
- results can only be improved by changing algorithm and/or assumptions,
- helps designing analytical proofs (reduces to linear combinations of inequalities),
  proofs can be engineered using numerics & symbolic computations!
- fast prototyping:
  before trying to prove your new FO method works; give PEP a try!
- step forward to "reproducible theory" (useful for reviewing, too ☺).

Difficulties:

- suffers from standard caveats of worst-case analyses,
  key is to find good assumptions/parametrization
- closed-form solutions might be involved.

Ongoing research directions, open questions:

- computer-assisted algorithmic design,
- adaptive & structure-exploiting methods,
- non-convex & non-Euclidean settings?
- best performing methods usually come with super weak guarantees (quasi-Newton, NL conjugate gradients, etc.): can we close the gap?

Concluding remarks

Performance estimation's philosophy

- ◇ numerically allows obtaining tight bounds (rigorous baselines),
- ◇ results can only be improved by changing algorithm and/or assumptions,
- ◇ helps designing analytical proofs (reduces to linear combinations of inequalities),
  proofs can be engineered using numerics & symbolic computations!
- ◇ fast prototyping:
  before trying to prove your new FO method works; give PEP a try!
- ◇ step forward to "reproducible theory" (useful for reviewing, too ☺).

Difficulties:

- ◇ suffers from standard caveats of worst-case analyses,
  key is to find good assumptions/parametrization
- ◇ closed-form solutions might be involved.

Ongoing research directions, open questions:

- ◇ computer-assisted algorithmic design,
- ◇ adaptive & structure-exploiting methods,
- ◇ non-convex & non-Euclidean settings?
- ◇ best performing methods usually come with super weak guarantees (quasi-Newton, NL conjugate gradients, etc.): can we close the gap?
- ◇ Higher order methods?

Take-home messages

Worst-cases are solutions to optimization problems.

Take-home messages

Worst-cases are solutions to optimization problems.

Sometimes, those optimization problems are tractable.

Take-home messages

Worst-cases are solutions to optimization problems.

Sometimes, those optimization problems are tractable.

Often tractable in convex optimization!

Any interest raised?

Main references:

- *"Smooth strongly convex interpolation and exact worst-case performance of first-order methods"* (with J. Hendrickx and F. Glineur),
- *"Exact worst-case performance of first-order methods for composite convex optimization"* (with J. Hendrickx and F. Glineur).
- *"Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions"* (with F. Bach)

Any interest raised?

Main references:

- *"Smooth strongly convex interpolation and exact worst-case performance of first-order methods"* (with J. Hendrickx and F. Glineur),
- *"Exact worst-case performance of first-order methods for composite convex optimization"* (with J. Hendrickx and F. Glineur).
- *"Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions"* (with F. Bach)

A few other recent directions (on my webpage):

- Stochastic methods

Any interest raised?

Main references:

- *"Smooth strongly convex interpolation and exact worst-case performance of first-order methods"* (with J. Hendrickx and F. Glineur),
- *"Exact worst-case performance of first-order methods for composite convex optimization"* (with J. Hendrickx and F. Glineur).
- *"Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions"* (with F. Bach)

A few other recent directions (on my webpage):

- Stochastic methods
- Monotone operators

Any interest raised?

Main references:

- *"Smooth strongly convex interpolation and exact worst-case performance of first-order methods"* (with J. Hendrickx and F. Glineur),
- *"Exact worst-case performance of first-order methods for composite convex optimization"* (with J. Hendrickx and F. Glineur).
- *"Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions"* (with F. Bach)

A few other recent directions (on my webpage):

- Stochastic methods
- Monotone operators
- Mirror descent, relative smoothness

Any interest raised?

Main references:

- *"Smooth strongly convex interpolation and exact worst-case performance of first-order methods"* (with J. Hendrickx and F. Glineur),
- *"Exact worst-case performance of first-order methods for composite convex optimization"* (with J. Hendrickx and F. Glineur).
- *"Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions"* (with F. Bach)

A few other recent directions (on my webpage):

- Stochastic methods
- Monotone operators
- Mirror descent, relative smoothness
- Attempts to the analysis of adaptive methods

Thanks! Questions?

www.di.ens.fr/$\sim$ataylor/

AdrienTaylor/Performance-Estimation-Toolbox on Github