Convergence rates in convex optimization
Beyond the worst-case with the help of geometry

Guillaume Garrigos

with Lorenzo Rosasco and Silvia Villa

École Normale Supérieure

Journées du GdR MOA/MIA - Bordeaux - 19 Oct 2017

Introduction

Setting: X Hilbert space, $f : X \to \mathbb{R} \cup \{+\infty\}$ convex l.s.c.
Problem: Minimize $f(x), \ x \in X$.
Tool: My favorite algorithm.

Setting: X Hilbert space, $f : X \to \mathbb{R} \cup \{+\infty\}$ convex l.s.c.
Problem: Minimize $f(x)$, $x \in X$.
Tool: My favorite algorithm.

As optimizers, we often face the same questions concerning the convergence of an algorithm:

- **(Qualitative result)** For the iterates $(x_n)_{n\in\mathbb{N}}$: weak, strong convergence?
- **(Quantitative result)** For the iterates and/or the values: sublinear $O\left(n^{-\alpha}\right)$ rates, linear $O\left(\varepsilon^{n}\right)$, superlinear ?

Introduction

Setting: X Hilbert space, $f : X \to \mathbb{R} \cup \{+\infty\}$ convex l.s.c.
Problem: Minimize $f(x)$, $x \in X$.
Tool: My favorite algorithm.

As optimizers, we often face the same questions concerning the convergence of an algorithm:

- **(Qualitative result)** For the iterates $(x_n)_{n\in\mathbb{N}}$: weak, strong convergence?
- **(Quantitative result)** For the iterates and/or the values: sublinear $O(n^{-\alpha})$ rates, linear $O(\varepsilon^n)$, superlinear ?

It depends on the algorithm and the assumptions made on f.

Introduction

Setting: X Hilbert space, $f : X \to \mathbb{R} \cup \{+\infty\}$ convex l.s.c.
Problem: Minimize $f(x)$, $x \in X$.
Tool: My favorite algorithm.

As optimizers, we often face the same questions concerning the convergence of an algorithm:

- **(Qualitative result)** For the iterates $(x_n)_{n\in\mathbb{N}}$: weak, strong convergence?
- **(Quantitative result)** For the iterates and/or the values: sublinear $O\left(n^{-\alpha}\right)$ rates, linear $O\left(\varepsilon^n\right)$, superlinear ?

It depends on the algorithm and the assumptions made on f.

Here we will essentially consider first order descent methods, and more simply the forward-backward method.

Setting: X Hilbert space, $f : X \to \mathbb{R} \cup \{+\infty\}$ convex l.s.c.
Problem: Minimize $f(x)$, $x \in X$.
Tool: My favorite algorithm.

As optimizers, we often face the same questions concerning the convergence of an algorithm:

- **(Qualitative result)** For the iterates $(x_n)_{n\in\mathbb{N}}$: weak, strong convergence?
- **(Quantitative result)** For the iterates and/or the values: sublinear $O\left(n^{-\alpha}\right)$ rates, linear $O\left(\varepsilon^n\right)$, superlinear ?

It depends on the algorithm and the assumptions made on f.

Here we will essentially consider first order descent methods, and more simply the forward-backward method.

Contents

Classic convergence results

Let $f = g + h$ be convex, with h L-Lipschitz smooth
Let $x_{n+1} = \text{prox}_{\lambda g}(x_n - \lambda \nabla h(x_n))$, $\lambda \in]0, 2/L[$.

Theorem (general convex case)

- argmin $f = \emptyset$: x_n diverges, no rates for $f(x_n) - \inf f$.
- argmin $f \neq \emptyset$: x_n weakly converges to $x_\infty \in$ argmin f, and $f(x_n) - \inf f = o\left(n^{-1}\right)$.

Classic convergence results

Let $f = g + h$ be convex, with h L-Lipschitz smooth
Let $x_{n+1} = \mathrm{prox}_{\lambda g}(x_n - \lambda \nabla h(x_n))$, $\lambda \in]0, 2/L[$.

Theorem (general convex case)

- $\mathrm{argmin}\, f = \emptyset$: x_n diverges, no rates for $f(x_n) - \inf f$.
- $\mathrm{argmin}\, f \neq \emptyset$: x_n weakly converges to $x_\infty \in \mathrm{argmin}\, f$, and $f(x_n) - \inf f = o\left(n^{-1}\right)$.

Theorem (strongly convex case)

Assume that f is strongly convex. Then x_n strongly converges to $x_\infty \in \mathrm{argmin}\, f$, and both iterates and values converge linearly.

Classic convergence results

Assume f to be convex and $(x_n)_{n\in\mathbb{N}}$ be generated by forward-backward.

function	values	iterates
$\operatorname{argmin} f = \emptyset$	$o(1)$	diverge
$\operatorname{argmin} f \neq \emptyset$	$o(n^{-1})$	weak convergence
s. convex	linear	linear

Classic convergence results

Assume f to be convex and $(x_n)_{n\in\mathbb{N}}$ be generated by forward-backward.

function	values	iterates
argmin $f = \emptyset$	$o(1)$	diverge
argmin $f \neq \emptyset$	$o(n^{-1})$	weak convergence
?	?	?
s. convex	linear	linear

Classic convergence results

Assume f to be convex and $(x_n)_{n\in\mathbb{N}}$ be generated by forward-backward.

function	values	iterates
$\operatorname{argmin} f = \emptyset$	$o(1)$	diverge
$\operatorname{argmin} f \neq \emptyset$	$o(n^{-1})$	weak convergence
?	?	?
?	linear	linear

Assume f to be convex and $(x_n)_{n\in\mathbb{N}}$ be generated by forward-backward.

function	values	iterates
$\operatorname{argmin} f = \emptyset$	$o(1)$	diverge
$\operatorname{argmin} f \neq \emptyset$	$o(n^{-1})$	weak convergence
?	?	?
?	linear	linear

$\longrightarrow$ Use geometry!

Known examples

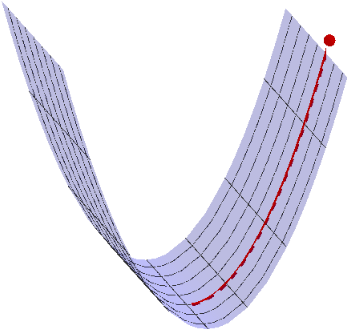
$A \in L(X, Y)$, $y \in Y$.

- $f(x) = \frac{1}{2}\|Ax - y\|^2$, $x_{n+1} = x_n - \tau A^*(Ax_n - y)$
  - If $R(A)$ is closed, linear convergence.

Known examples

$A \in L(X, Y)$, $y \in Y$.

- $f(x) = \frac{1}{2}\|Ax - y\|^2$, $x_{n+1} = x_n - \tau A^*(Ax_n - y)$
  - If $R(A)$ is closed, linear convergence.



Known examples

$A \in L(X, Y)$, $y \in Y$.

- $f(x) = \frac{1}{2}\|Ax - y\|^2$, $x_{n+1} = x_n - \tau A^*(Ax_n - y)$
  - If $R(A)$ is closed, linear convergence.
  - Else, strong convergence for iterates, arbitrarily slow.

$A \in L(X,Y)$, $y \in Y$.

- $f(x) = \frac{1}{2}\|Ax - y\|^2$, $x_{n+1} = x_n - \tau A^*(Ax_n - y)$
  - If $R(A)$ is closed, linear convergence.
  - Else, strong convergence for iterates, arbitrarily slow.
- $f(x) = \alpha\|x\|_1 + \frac{1}{2}\|Ax - y\|^2$, $x_{n+1} = \mathcal{S}_{\alpha\tau}\left(x_n - \tau A^*(Ax_n - y)\right)$
  - In $X = \mathbb{R}^N$, the convergence is linear.[1]
  - In $X = \ell^2(\mathbb{N})$, ISTA converges strongly[2]. Linear rates can also be obtained under some conditions[3]. In fact not necessary[4].

[1] Bolte, Nguyen, Peypouquet, Suter (2015), based on Li (2012)
[2] Daubechies, Defrise, DeMol (2004)
[3] Bredies, Lorenz (2008)
[4] End of this talk

Known examples

$A \in L(X, Y)$, $y \in Y$.

- $f(x) = \frac{1}{2}\|Ax - y\|^2$, $x_{n+1} = x_n - \tau A^*(Ax_n - y)$
  - If $R(A)$ is closed, linear convergence.
  - Else, strong convergence for iterates, arbitrarily slow.
- $f(x) = \alpha\|x\|_1 + \frac{1}{2}\|Ax - y\|^2$, $x_{n+1} = \mathcal{S}_{\alpha\tau}\left(x_n - \tau A^*(Ax_n - y)\right)$
  - In $X = \mathbb{R}^N$, the convergence is linear.[1]
  - In $X = \ell^2(\mathbb{N})$, ISTA converges strongly[2]. Linear rates can also be obtained under some conditions[3]. In fact not necessary[4].
- Gap between theory and practice.

[1] Bolte, Nguyen, Peypouquet, Suter (2015), based on Li (2012)
[2] Daubechies, Defrise, DeMol (2004)
[3] Bredies, Lorenz (2008)
[4] End of this talk

Contents

Let $p \geq 1$ and $\Omega \subset X$ and arbitrary set.

Definition

We say that f is p-conditioned on Ω if $\exists \gamma_\Omega > 0$ such that

$$\forall x \in \Omega, \ \frac{\gamma_\Omega}{p} \mathrm{dist}\,(x, \operatorname{argmin} f)^p \leq f(x) - \inf f.$$

Conditioned and Lojasiewicz functions

Let $p \geq 1$ and $\Omega \subset X$ and arbitrary set.

Definition

We say that f is p-conditioned on Ω if $\exists \gamma_\Omega > 0$ such that

$$\forall x \in \Omega, \ \frac{\gamma_\Omega}{p} \mathrm{dist}\,(x, \operatorname{argmin} f)^p \leq f(x) - \inf f.$$

- The exponent p governs the local geometry of f, and then the rates of convergence. Easy to get.

Conditioned and Lojasiewicz functions

Let $p \geq 1$ and $\Omega \subset X$ and arbitrary set.

Definition

We say that f is p-conditioned on Ω if $\exists \gamma_\Omega > 0$ such that

$$\forall x \in \Omega, \;\; \frac{\gamma_\Omega}{p} \mathrm{dist}\,(x, \operatorname{argmin} f)^p \leq f(x) - \inf f.$$

- The exponent p governs the local geometry of f, and then the rates of convergence. Easy to get.
- γ_Ω governs the constant in the rates. Hard to estimate properly.

[1] Bolte, Nguyen, Peypouquet, Suter, 2015 - Garrigos, Rosasco , Villa, 2016.

Conditioned and Lojasiewicz functions

Let $p \geq 1$ and $\Omega \subset X$ and arbitrary set.

Definition

We say that f is p-conditioned on Ω if $\exists \gamma_\Omega > 0$ such that

$$\forall x \in \Omega, \ \frac{\gamma_\Omega}{p} \mathrm{dist}\,(x, \operatorname{argmin} f)^p \leq f(x) - \inf f.$$

- The exponent p governs the local geometry of f, and then the rates of convergence. Easy to get.
- γ_Ω governs the constant in the rates. Hard to estimate properly.
- "Equivalent" to Lojasiewicz inequality/metric subregularity[1].

[1] Bolte, Nguyen, Peypouquet, Suter, 2015 - Garrigos, Rosasco , Villa, 2016.

Identifying the geometry: Some examples

- strongly convex functions are 2-conditioned on X, $\gamma_X = \gamma$

Identifying the geometry: Some examples

- strongly convex functions are 2-conditioned on X, $\gamma_X = \gamma$
- $f(x) = \frac{1}{2}\|Ax - y\|^2$
  - If $R(A)$ is closed, f is 2-conditioned on X, $\gamma_X = \sigma^*_{min}(A^*A)$.

Identifying the geometry: Some examples

- strongly convex functions are 2-conditioned on X, $\gamma_X = \gamma$
- $f(x) = \frac{1}{2}\|Ax - y\|^2$
  - If $R(A)$ is closed, f is 2-conditioned on X, $\gamma_X = \sigma^*_{min}(A^*A)$.
  - Else, complicated (see later).

- strongly convex functions are 2-conditioned on X, $\gamma_X = \gamma$
- $f(x) = \frac{1}{2}\|Ax - y\|^2$
  - If $R(A)$ is closed, f is 2-conditioned on X, $\gamma_X = \sigma^*_{min}(A^*A)$.
  - Else, complicated (see later).
- In $\mathbb{R}^N$, convex polynomial by parts functions are p-conditioned[1] on sublevel sets, with $p = 1 + (d-1)^N$, but $\gamma_{[f \leq r]}$ unknown. Example: $f(x) = \alpha\|x\|_1 + \frac{1}{2}\|Ax - y\|^2$.

[1] Yang, 2009 + Li, 2012

- strongly convex functions are 2-conditioned on X, $\gamma_X = \gamma$
- $f(x) = \frac{1}{2}\|Ax - y\|^2$
  - If $R(A)$ is closed, f is 2-conditioned on X, $\gamma_X = \sigma^*_{min}(A^*A)$.
  - Else, complicated (see later).
- In $\mathbb{R}^N$, convex polynomial by parts functions are p-conditioned[1] on sublevel sets, with $p = 1 + (d-1)^N$, but $\gamma_{[f \leq r]}$ unknown. Example: $f(x) = \alpha\|x\|_1 + \frac{1}{2}\|Ax - y\|^2$.
- Almost any simple function used in practice: $\|x\|^p_\alpha$, KL divergence, etc...

[1] Yang, 2009 + Li, 2012

Identifying the geometry: Some examples

- strongly convex functions are 2-conditioned on X, $\gamma_X = \gamma$
- $f(x) = \frac{1}{2}\|Ax - y\|^2$
  - If $R(A)$ is closed, f is 2-conditioned on X, $\gamma_X = \sigma^*_{min}(A^*A)$.
  - Else, complicated (see later).
- In $\mathbb{R}^N$, convex polynomial by parts functions are p-conditioned[1] on sublevel sets, with $p = 1 + (d-1)^N$, but $\gamma_{[f \leq r]}$ unknown. Example: $f(x) = \alpha\|x\|_1 + \frac{1}{2}\|Ax - y\|^2$.
- Almost any simple function used in practice: $\|x\|^p_\alpha$, KL divergence, etc...
- semi-algebraic functions are conditioned around minimizers[2]. p and γ unknown.

[1] Yang, 2009 + Li, 2012
[2] Bolte, Daniilidis, Lewis, Shiota, 2007

Identifying the geometry: two rules

Theorem: Sum rule[1]

Assume that f_1 and f_2 are respectively p_1 and p_2-conditioned, up to linear perturbations, on $\Omega \subset X$. Then, under some qualification condition, $f_1 + f_2$ is p-conditioned on Ω with $p = \max\{p_1, p_2\}$.

Theorem: Composition with linear operator (closed range)[1]

Assume that f is p-conditioned and smooth, up to linear perturbations, on $\Omega \subset X$. Then, under some qualification conditions, $f \circ A$ is p-conditioned on $A^{-1}\Omega$.

[1] Lewis, Drusvyatskiy (2016) for $p = 2$; G., Rosasco, Villa (2016) for $p \geq 1$.

Theorem: Sum rule[1]

Assume that f_1 and f_2 are respectively p_1 and p_2-conditioned, up to linear perturbations, on $\Omega \subset X$. Then, under some qualification condition, $f_1 + f_2$ is p-conditioned on Ω with $p = \max\{p_1, p_2\}$.

Theorem: Composition with linear operator (closed range)[1]

Assume that f is p-conditioned and smooth, up to linear perturbations, on $\Omega \subset X$. Then, under some qualification conditions, $f \circ A$ is p-conditioned on $A^{-1}\Omega$.

Not always true without QC! See $\|M\|_* + \|\mathcal{A}M - \mathcal{Y}\|^2$.

[1] Lewis, Drusvyatskiy (2016) for $p = 2$; G., Rosasco, Villa (2016) for $p \geq 1$.

Contents

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

Let $(x_n)_{n\in\mathbb{N}}$ be generated by the Forward-Backward, and suppose

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Then x_n converges strongly to a minimizer $x^\dagger$ of f. Moreover, $\forall n \in \mathbb{N}$:

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

Let $(x_n)_{n\in\mathbb{N}}$ be generated by the Forward-Backward, and suppose

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Then x_n converges strongly to a minimizer $x^\dagger$ of f. Moreover, $\forall n \in \mathbb{N}$:

1. if $p = 2$, linear convergence with $\varepsilon \in]0,1[$, $C > 0$

$$f(x_{n+1}) - \inf f \leq \varepsilon(f(x_n) - \inf f) \text{ and } \|x_n - x^\dagger\| \leq C\sqrt{\varepsilon}^n,$$

Exploiting the geometry: Convergence result

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

Let $(x_n)_{n\in\mathbb{N}}$ be generated by the Forward-Backward, and suppose

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Then x_n converges strongly to a minimizer $x^\dagger$ of f. Moreover, $\forall n \in \mathbb{N}$:

1. if $p = 2$, linear convergence with $\varepsilon \in]0,1[$, $C > 0$

$$f(x_{n+1}) - \inf f \leq \varepsilon(f(x_n) - \inf f) \text{ and } \|x_n - x^\dagger\| \leq C\sqrt{\varepsilon}^n,$$

2. if $p > 2$, sublinear convergence with $C_1, C_2 > 0$

$$f(x_n) - \inf f \leq C_1 n^{\frac{-p}{p-2}} \text{ and } \|x_n - x^\dagger\| \leq C_2 n^{\frac{-1}{p-2}}.$$

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

Let $(x_n)_{n\in\mathbb{N}}$ be generated by the Forward-Backward, and suppose

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Then x_n converges strongly to a minimizer $x^\dagger$ of f. Moreover, $\forall n \in \mathbb{N}$:

1. if $p = 2$, linear convergence with $\varepsilon \in]0,1[$, $C > 0$

$$f(x_{n+1}) - \inf f \leq \varepsilon(f(x_n) - \inf f) \text{ and } \|x_n - x^\dagger\| \leq C\sqrt{\varepsilon}^n,$$

2. if $p > 2$, sublinear convergence with $C_1, C_2 > 0$

$$f(x_n) - \inf f \leq C_1 n^{\frac{-p}{p-2}} \text{ and } \|x_n - x^\dagger\| \leq C_2 n^{\frac{-1}{p-2}}.$$

NB: All the constants depend on $(L, \lambda, p, \gamma_{f,\Omega}, f(x^0) - \inf f)$.

Exploiting the geometry: convergence result

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Then $p = 2$ gives linear rates, $p > 2$ sublinear rates.

Some remarks on the convergence result:

Exploiting the geometry: convergence result

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Then $p = 2$ gives linear rates, $p > 2$ sublinear rates.

Some remarks on the convergence result:

- These rates are *optimal* (see $f(x) = \|x\|^p$).

Exploiting the geometry: convergence result

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Then $p = 2$ gives linear rates, $p > 2$ sublinear rates.

Some remarks on the convergence result:

- These rates are *optimal* (see $f(x) = \|x\|^p$).
- Rates involve a generalized *condition number* $\kappa \propto L/\gamma_{f,\Omega}$.
  For $p = 2$ there is $\varepsilon = \kappa/(\kappa + 1)$.

Exploiting the geometry: convergence result

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n \in \mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Then $p = 2$ gives linear rates, $p > 2$ sublinear rates.

Some remarks on the convergence result:

- These rates are *optimal* (see $f(x) = \|x\|^p$).
- Rates involve a generalized *condition number* $\kappa \propto L/\gamma_{f,\Omega}$.
  For $p = 2$ there is $\varepsilon = \kappa/(\kappa + 1)$.
- These results extends to the nonconvex setting.

Exploiting the geometry: convergence result

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Then $p = 2$ gives linear rates, $p > 2$ sublinear rates.

Some remarks on the convergence result:

- These rates are *optimal* (see $f(x) = \|x\|^p$).
- Rates involve a generalized *condition number* $\kappa \propto L/\gamma_{f,\Omega}$.
  For $p = 2$ there is $\varepsilon = \kappa/(\kappa + 1)$.
- These results extends to the nonconvex setting.
- These results extends to general first-order descent methods.

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n \in \mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

- Clarify the local vs global rates.

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

• Clarify the local vs global rates.
$\exists(\delta, r) \in]0, +\infty[^2$, f is p-conditioned on $\Omega := B(\bar{x}, \delta) \cap [f - \inf \leq r]$.

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

• Clarify the local vs global rates.
$\exists(\delta, r) \in]0, +\infty[^2$, f is p-conditioned on $\Omega := B(\bar{x}, \delta) \cap [f - \inf \leq r]$.
Fejer + descent $\Rightarrow \exists N \in \mathbb{N}, \forall n \geq N, \quad x^n \in \Omega \Rightarrow$ Local rates.

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

• Clarify the local vs global rates.
$\forall(\delta, r) \in]0, +\infty[^2$, f is p-conditioned on $\Omega := B(\bar{x}, \delta) \cap [f - \inf \leq r]$.

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

• Clarify the local vs global rates.
$\forall(\delta, r) \in]0, +\infty[^2$, f is p-conditioned on $\Omega := B(\bar{x}, \delta) \cap [f - \inf \leq r]$.
Fejer + descent $\Rightarrow \forall n \geq 0, x^n \in \Omega, \Rightarrow$ Global rates.

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

- Clarify the local vs global rates.
- Some functions have nonlocal geometry Ω : $f(x) = \|Ax - y\|^2$.

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

- Clarify the local vs global rates.
- Some functions have nonlocal geometry Ω : $f(x) = \|Ax - y\|^2$.
    - If $\operatorname{Im} A$ not closed, Haraux and Jendoubi show that no conditioning hold on $B(\bar{x}, \delta)$.

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

- Clarify the local vs global rates.
- Some functions have nonlocal geometry Ω : $f(x) = \|Ax - y\|^2$.
  - If $\mathrm{Im}\, A$ not closed, Haraux and Jendoubi show that no conditioning hold on $B(\bar{x}, \delta)$.
  - We prove that conditioning holds on "Sobolev" spaces.

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n \in \mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

- Clarify the local vs global rates.
- Some functions have nonlocal geometry Ω : $f(x) = \|Ax - y\|^2$.
- We can restrict to low-dimensional sets.

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

- Clarify the local vs global rates.
- Some functions have nonlocal geometry Ω : $f(x) = \|Ax - y\|^2$.
- We can restrict to low-dimensional sets.

If $f = g + h$ with h smooth and g partially smooth + QC, then $\exists N \in N$, $\forall n \geq N$, $x^n \in \mathcal{M}$ (identification of active manifold)

On the localization/geometry trade-off

Theorem (G., Rosasco, Villa, 2016) & (Frankel, G., Peypouquet, 2014)

- ***(Localization)*** $(x_n)_{n\in\mathbb{N}} \subset \Omega$,
- ***(Geometry)*** f is p-conditioned on Ω.

Localization hypothesis seems a trick. And why general $\Omega \subset X$?

- Clarify the local vs global rates.
- Some functions have nonlocal geometry Ω : $f(x) = \|Ax - y\|^2$.
- We can restrict to low-dimensional sets.

If $f = g + h$ with h smooth and g partially smooth + QC, then
$\exists N \in N$, $\forall n \geq N$, $x^n \in \mathcal{M}$ (identification of active manifold)
$\rightarrow$ conditioning on $\mathcal{M}$ is enough, no need for strong convexity.

Updated results

function	values	iterates
argmin $f = \emptyset$	$o(1)$	diverge
argmin $f \neq \emptyset$	$o(n^{-1})$	weak convergence
geometry, $p > 2$	$O\left(n^{\frac{-p}{p-2}}\right)$	$O\left(n^{\frac{-1}{p-2}}\right)$
geometry, $p = 2$	linear	linear
geometry, $1 < p < 2$	superlinear	superlinear
geometry, $p = 1$	finite	finite

Updated results

function	values	iterates
argmin $f = \emptyset$	$o(1)$	diverge
argmin $f \neq \emptyset$	$o(n^{-1})$	weak convergence
geometry, $p > 2$	$O\left(n^{\frac{-p}{p-2}}\right)$	$O\left(n^{\frac{-1}{p-2}}\right)$
geometry, $p = 2$	linear	linear
geometry, $1 < p < 2$	superlinear	superlinear
geometry, $p = 1$	finite	finite

- We have a spectra covering "almost" all convex functions in finite dimensions[1] .

[1] Bolte, Daniilidis, Ley, Mazet - 2010

Updated results

function	values	iterates
argmin $f = \emptyset$	$o(1)$	diverge
argmin $f \neq \emptyset$	$o(n^{-1})$	weak convergence
geometry, $p > 2$	$O\left(n^{\frac{-p}{p-2}}\right)$	$O\left(n^{\frac{-1}{p-2}}\right)$
geometry, $p = 2$	linear	linear
geometry, $1 < p < 2$	superlinear	superlinear
geometry, $p = 1$	finite	finite

- We have a spectra covering "almost" all convex functions in finite dimensions[1] .
- The hypothesis to get linear rates is minimal

[1] Bolte, Daniilidis, Ley, Mazet - 2010

Updated results

function	values	iterates
argmin $f = \emptyset$	$o(1)$	diverge
argmin $f \neq \emptyset$	$o(n^{-1})$	weak convergence
geometry, $p > 2$	$O\left(n^{\frac{-p}{p-2}}\right)$	$O\left(n^{\frac{-1}{p-2}}\right)$
geometry, $p = 2$	linear	linear
geometry, $1 < p < 2$	superlinear	superlinear
geometry, $p = 1$	finite	finite

Proposition

If linear rates hold on Ω:

$$(\exists \varepsilon \in]0,1[)(\forall x \in \Omega) \quad \operatorname{dist}(FB(x), \operatorname{argmin} f) \leq \varepsilon \operatorname{dist}(x, \operatorname{argmin} f),$$

then f is 2-conditioned on Ω.

Updated results

function	values	iterates
argmin $f = \emptyset$	$o(1)$	diverge
argmin $f \neq \emptyset$	$o(n^{-1})$	weak convergence
geometry, $p > 2$	$O\left(n^{\frac{-p}{p-2}}\right)$	$O\left(n^{\frac{-1}{p-2}}\right)$
geometry, $p = 2$	linear	linear
geometry, $1 < p < 2$	superlinear	superlinear
geometry, $p = 1$	finite	finite

- We have a spectra covering "almost" all convex functions in finite dimensions[1] .
- The hypothesis to get linear rates is minimal
- Up to now, the infinite dimensional setting is less understood.

[1] Bolte, Daniilidis, Ley, Mazet - 2010

Contents

Least squares: $f(x) = \frac{1}{2}\|Ax - y\|^2$

Assume that $R(A)$ is not closed, and $y \in \operatorname{dom} A^\dagger$.
The FB method becomes $x_{n+1} = x_n - \lambda A^*(Ax_n - y)$, $x_0 = 0$.

Least squares: $f(x) = \frac{1}{2}\|Ax - y\|^2$

Assume that $R(A)$ is not closed, and $y \in \operatorname{dom} A^\dagger$.
The FB method becomes $x_{n+1} = x_n - \lambda A^*(Ax_n - y)$, $x_0 = 0$.

x_n converges to $x^\dagger := A^\dagger y$. But how fast?
$\rightarrow$ Old answer: it depends on the regularity of $x^\dagger$.

Least squares: $f(x) = \frac{1}{2}\|Ax - y\|^2$

Assume that $R(A)$ is not closed, and $y \in \mathrm{dom}\, A^\dagger$.
The FB method becomes $x_{n+1} = x_n - \lambda A^*(Ax_n - y)$, $x_0 = 0$.

x_n converges to $x^\dagger := A^\dagger y$. But how fast?
$\rightarrow$ Old answer: it depends on the regularity of $x^\dagger$.

In inverse problems, the spaces $R(A^*A^\mu)$ play the role of Sobolev in L^2.

Example: Sobolev regularity

If $X = Y = L^2([0, 2\pi])$ and A is the integration operator, then

$$R(A^*A^\mu) = H^{2\mu}([0, 2\pi]).$$

Least squares: Convergence analysis

Theorem: Geometry on Sobolev spaces

The least squares f is p-conditioned on each affine space $x^\dagger + R(A^*A^\mu)$, with the exponent $p = 2 + \mu^{-1}$.

Fact: if $x^\dagger \in R(A^*A^\mu)$ and $x_0 = 0$, then $(x_n)_{n\in\mathbb{N}} \subset x^\dagger + R(A^*A^\mu)$.

Theorem: Convergence for Landweber's algorithm

If $x_0 = 0$, and $x^\dagger \in R(A^*A^\mu)$, then the convergence is sublinear:

$$f(x_n) - \inf f = O\left(n^{-(1+2\mu)}\right) \text{ and } \|x_n - x^\dagger\| = O\left(n^{-\mu}\right).$$

NB: the exponent $p = 2 + \mu^{-1}$ and the rates are **tight**.

Least squares: what if $\operatorname{argmin} \|Ax - y\|^2 = \emptyset$?

It might be that $x^\dagger = A^\dagger y$ doesn't exist...

Least squares: what if $\operatorname{argmin} \|Ax - y\|^2 = \emptyset$?

It might be that $x^\dagger = A^\dagger y$ doesn't exist...

Typically in learning we look for a function in $L^2(\mathcal{X} \times \mathcal{Y}, \rho)$
But in practice we work in a RKHS $X \subset L^2$

It might be that $x^\dagger = A^\dagger y$ doesn't exist...

Typically in learning we look for a function in $L^2(\mathcal{X} \times \mathcal{Y}, \rho)$
But in practice we work in a RKHS $X \subset L^2$

Even if f has no minimizers, we still want to estimate $f(x^n) - \inf f \to 0$
It will depend on how far the solution is from X.

Least squares: what if $\operatorname{argmin} \|Ax - y\|^2 = \emptyset$?

It might be that $x^\dagger = A^\dagger y$ doesn't exist...

Typically in learning we look for a function in $L^2(\mathcal{X} \times \mathcal{Y}, \rho)$
But in practice we work in a RKHS $X \subset L^2$

Even if f has no minimizers, we still want to estimate $f(x^n) - \inf f \to 0$
It will depend on how far the solution is from X.

We look at how regular is $y^\dagger := \mathrm{proj}\,(y, \overline{\operatorname{Im} A})$ within $\overline{\operatorname{Im} A} \subset Y$.

Least squares but no minimizers: Convergence analysis

Theorem: Geometry on Sobolev spaces (w.r.t. data space Y)

The least squares f is "p-conditioned" on each affine space

$$A^{-1}\left(y^\dagger + R(AA^{*\nu})\right),\ \nu > 0$$

with the exponent $p = 2 + (\nu - 1/2)^{-1}$.

Fact: if $\nu < 1/2$ then $p < 0$!! f behaves like $\frac{1}{t^{|p|}}$.

Theorem: Convergence for Landweber's algorithm

If $x_0 = 0$, and $y^\dagger \in R(AA^{*\nu})$, then the convergence is sublinear:

$$f(x_n) - \inf f = O\left(n^{-2\nu}\right).$$

Updated results

Assume f to be convex and $(x_n)_{n\in\mathbb{N}}$ be generated by a first-order descent method.

function	values	iterates
argmin $f = \emptyset$	$o(1)$	diverge
geometry, $p < 0$	$O\left(n^{\frac{-p}{p-2}}\right)$	diverge
argmin $f \neq \emptyset$	$o(n^{-1})$	weak convergence
geometry, $p > 2$	$O\left(n^{\frac{-p}{p-2}}\right)$	$O\left(n^{\frac{-1}{p-2}}\right)$
geometry, $p = 2$	linear	linear
geometry, $1 < p < 2$	superlinear	superlinear
geometry, $p = 1$	finite	finite

Contents

Lasso in Hilbert spaces

Consider the Lasso in $\ell^2(\mathbb{N})$

$$f(x) = \alpha\|x\|_1 + \frac{1}{2}\|Ax - y\|^2$$

How fast do converge ISTA? $O(1/n)$? linearly?

Consider the Lasso in $\ell^2(\mathbb{N})$

$$f(x) = \alpha\|x\|_1 + \frac{1}{2}\|Ax - y\|^2$$

How fast do converge ISTA? $O(1/n)$? linearly?

- linear rates if A is injective on finite supports
- linear rates if qualification condition holds

Lasso in Hilbert spaces

Consider the Lasso in $\ell^2(\mathbb{N})$

$$f(x) = \alpha\|x\|_1 + \frac{1}{2}\|Ax - y\|^2$$

How fast do converge ISTA? $O(1/n)$? linearly?

- linear rates if A is injective on finite supports
- linear rates if qualification condition holds

Theorem (G., Rosasco, Villa - 2017)

There exists Ω such that $(x_n)_{n\in\mathbb{N}} \subset \Omega$ and f is 2-conditioned on Ω. So ISTA always converge linearly.

Lasso in Hilbert spaces

Consider the Lasso in $\ell^2(\mathbb{N})$

$$f(x) = \alpha \|x\|_1 + \frac{1}{2}\|Ax - y\|^2$$

How fast do converge ISTA? $O(1/n)$? linearly?

- linear rates if A is injective on finite supports
- linear rates if qualification condition holds

Theorem (G., Rosasco, Villa - 2017)

There exists Ω such that $(x_n)_{n\in\mathbb{N}} \subset \Omega$ and f is 2-conditioned on Ω. So ISTA always converge linearly.

Similar result by replacing $\|\cdot\|_1$ with $\|\cdot\|_1 + \|\cdot\|_p^p$.

Conclusion

If you had to remember ONE thing

You have a descent-related (dissipative?) algorithm?
Strong convexity gives you strong convergence and better rates?

Try to use the 2-conditioning:

$$\gamma \operatorname{dist}(x, \operatorname{argmin} f)^2 \leq f(x) - \inf f$$

$\longrightarrow$ It should give the same results than strong convexity
$\longrightarrow$ It applies to a way more general class of functions (actually super sharp for linear rates)

Conclusion/Discussion

- Structural results allow a practical identification of geometry.

Conclusion/Discussion

- Structural results allow a practical identification of geometry.
- Geometry sheds a new light on *a priori* unrelated results.

Conclusion/Discussion

- Structural results allow a practical identification of geometry.
- Geometry sheds a new light on *a priori* unrelated results.
- Quantitative characterization of the geometry in the nonconvex case is an active topic. E.g.: $f(w) = \sum \ell(\langle x_i, w \rangle - y_i)$.

Conclusion/Discussion

- Structural results allow a practical identification of geometry.
- Geometry sheds a new light on *a priori* unrelated results.
- Quantitative characterization of the geometry in the nonconvex case is an active topic. E.g.: $f(w) = \sum \ell(\langle x_i, w\rangle - y_i)$.
- Descent methods very well understood. Holds for general first-order descent methods
  1. **(descent)** $a\|x_{n+1} - x_n\|^2 \leq f(x_n) - f(x_{n+1})$
  2. **(1st order)** $b\|\partial f(x_{n+1})\|_- \leq \|x_{n+1} - x_n\|$

  Allows even more structured methods (decomposition by blocs), or variants (variable metric, inexact computations)

Conclusion/Discussion

- Structural results allow a practical identification of geometry.
- Geometry sheds a new light on *a priori* unrelated results.
- Quantitative characterization of the geometry in the nonconvex case is an active topic. E.g.: $f(w) = \sum \ell(\langle x_i, w \rangle - y_i)$.
- Descent methods very well understood. Holds for general first-order descent methods
- Recently: application to stochastic gradient methods.

Conclusion/Discussion

- Structural results allow a practical identification of geometry.
- Geometry sheds a new light on *a priori* unrelated results.
- Quantitative characterization of the geometry in the nonconvex case is an active topic. E.g.: $f(w) = \sum \ell(\langle x_i, w \rangle - y_i)$.
- Descent methods very well understood. Holds for general first-order descent methods
- Recently: application to stochastic gradient methods.
- Geometry is a powerful tool not only for rates, but also for regularization! (see Silvia's talk)

Conclusion/Discussion

- Structural results allow a practical identification of geometry.
- Geometry sheds a new light on *a priori* unrelated results.
- Quantitative characterization of the geometry in the nonconvex case is an active topic. E.g.: $f(w) = \sum \ell(\langle x_i, w \rangle - y_i)$.
- Descent methods very well understood. Holds for general first-order descent methods
- Recently: application to stochastic gradient methods.
- Geometry is a powerful tool not only for rates, but also for regularization! (see Silvia's talk)
- Can inertial methods benefit from this analysis? Are they adaptive?

Thanks for your attention !