Métodos inerciais para resolver problemas de otimização convexa

Majela Pentón
UFBA

Novembro, 2024

Minimização convexa e aplicações

Métodos do gradiente e do subgradiente

Métodos inerciais

Minimização convexa

$$\min f(x)$$
$$\text{s.a. } x \in \mathbb{R}^n$$

- $f : \mathbb{R}^n \to \mathbb{R}$ função convexa

Minimização convexa

$$\begin{aligned} &\min f(x) \\ &\text{s.a. } x \in \mathbb{R}^n \end{aligned}$$

- $f : \mathbb{R}^n \to \mathbb{R}$ função convexa
- $x^\star$ solução do problema
- $f^\star = f(x^\star)$ valor ótimo

Aplicações

Aprendizado de máquina (machine learning)

Tem como objetivo principal tomar decisões ou fazer predições baseado em dados

Aplicações

Aprendizado de máquina (machine learning)

Tem como objetivo principal tomar decisões ou fazer predições baseado em dados

- Reconhecimento facial
- Processamento de linguagem (tradutor de Google)
- Sistemas de recomendação (Netflix, Amazon, ...)
- Predição de tumores
- Estimar preços
- Decidir se um empréstimo será aprovado ou não

Mínimos quadrados ℓ_1-regularizado (Lasso)

O objetivo do problema é encontrar uma solução esparsa do sistema de equações $Ax = b$.

$$\begin{aligned} &\min \frac{1}{2} \|Ax - b\|_2^2 + \mu \|x\|_1 \\ &\text{s.a. } x \in \mathbb{R}^n \end{aligned}$$

Mínimos quadrados ℓ_1-regularizado (Lasso)

O objetivo do problema é encontrar uma solução esparsa do sistema de equações $Ax = b$.

$$\begin{aligned} &\min \frac{1}{2} \|Ax - b\|_2^2 + \mu \|x\|_1 \\ &\text{s.a. } x \in \mathbb{R}^n \end{aligned}$$

- $A \in \mathbb{R}^{m \times n}$
- $b \in \mathbb{R}^m$
- $\mu > 0$

Aplicações

Processamento de imagens e sinais

- Remoção de ruído de imagens (modelo total variation)
- Remoção de borrados de imagens (deblurring)
- Compressed sensing

Total variation

Para reconstruir uma imagem $u \in \mathbb{R}^{m \times n}$ a partir de uma imagem borrada e/ou ruidosa $b \in \mathbb{R}^{m \times n}$, que descreve os dados obtidos, como solução do problema:

$$\min_{u \in \mathbb{R}^{m \times n}} \mu TV(u) + \frac{1}{2} \|u - b\|_F^2$$

Total variation

Para reconstruir uma imagem $u \in \mathbb{R}^{m \times n}$ a partir de uma imagem borrada e/ou ruidosa $b \in \mathbb{R}^{m \times n}$, que descreve os dados obtidos, como solução do problema:

$$\min_{u \in \mathbb{R}^{m \times n}} \mu TV(u) + \frac{1}{2} \|u - b\|_F^2$$

$$\min_{u \in \mathbb{R}^{m \times n}} \mu TV(u) + \frac{1}{2} \|u - b\|_F^2$$

- $\mu > 0$
- $\|\cdot\|_F$ é a norma de Frobenius para matrizes
- TV é a norma de variação total

$$TV(u) = \|\nabla_1 u\|_1 + \|\nabla_2 u\|_1$$

onde

$$(\nabla_1 u)_{ij} = u_{i+1,j} - u_{i,j} \quad \text{e} \quad (\nabla_2 u)_{ij} = u_{i,j+1} - u_{i,j}$$

Minimização convexa

- Todo mínimo local é global

Minimização convexa

- Todo mínimo local é global
- Se a função é diferenciável, a condição necessária de primeira ordem é suficiente.
  Condição necessária de primeira ordem: Se f é uma função diferenciável em $\mathbb{R}^n$ e $x^\star$ é um mínimo local de f então

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

Método do gradiente

- $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa diferenciável.

Método do gradiente

- $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa diferenciável.

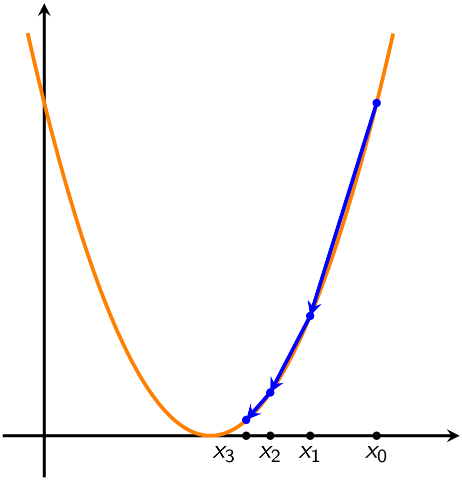
Gera uma sequência $(f(x_k))_{k\in\mathbb{N}}$ tal que $f(x_{k+1}) \leq f(x_k)$

Método do gradiente

- $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa diferenciável.

Gera uma sequência $(f(x_k))_{k\in\mathbb{N}}$ tal que $f(x_{k+1}) \leq f(x_k)$

- Escolher um ponto inicial $x_0 \in \mathbb{R}^n$
- Iterar: $x_{k+1} = x_k - \alpha\nabla f(x_k)$, $k = 0, 1, \ldots$
- $\alpha > 0$ é o **comprimento de passo** tal que $f(x_{k+1}) \leq f(x_k)$.


x_3
x_2
x_1
x_0

Convergência

$f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, diferenciável e ∇f é L-Lipschitz contínuo (com $L > 0$) em $\mathbb{R}^n$:

- Se $(x_k)_{k\in\mathbb{N}}$ é gerada pelo método do gradiente com passo $\alpha = 1/L$, então a sequência converge a uma solução global.

Convergência

$f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, diferenciável e ∇f é L-Lipschitz contínuo (com $L > 0$) em $\mathbb{R}^n$:

- Se $(x_k)_{k \in \mathbb{N}}$ é gerada pelo método do gradiente com passo $\alpha = 1/L$, então a sequência converge a uma solução global.

ϵ-solução

Dado $\epsilon > 0$, $\overline{x}$ é uma ϵ-solução se $f(\overline{x}) - f^\star \leq \epsilon$

Convergência

$f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, diferenciável e ∇f é L-Lipschitz contínuo (com $L > 0$) em $\mathbb{R}^n$:

- Se $(x_k)_{k\in\mathbb{N}}$ é gerada pelo método do gradiente com passo $\alpha = 1/L$, então a sequência converge a uma solução global.

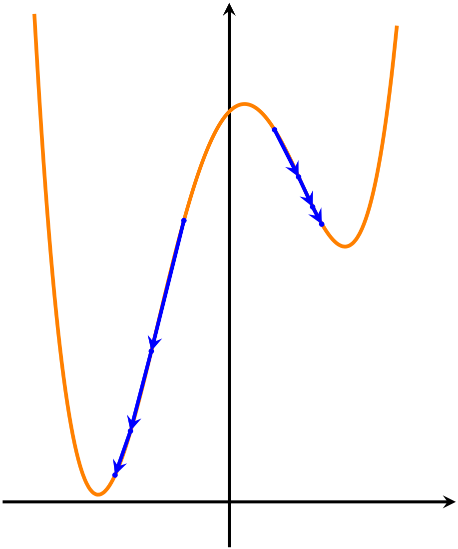
ϵ-solução

Dado $\epsilon > 0$, $\overline{x}$ é uma ϵ-solução se $f(\overline{x}) - f^\star \leq \epsilon$

Taxa de convergência

- O método do gradiente tem **taxa de convergência** $\mathcal{O}(1/k)$. Ou seja, o # de iterações para obter uma ϵ-**solução** é $\mathcal{O}(1/\epsilon)$.

- Se f não é convexa, o método do gradiente pode não convergir para um mínimo global da função.



Método do gradiente

Vantagens:

- ideia simples
- fácil implementação
- usualmente as iterações são baratas

Método do gradiente

Vantagens:

- ideia simples
- fácil implementação
- usualmente as iterações são baratas

Desvantagens:

- não pode ser aplicado a funções não diferenciáveis
  - $\frac{1}{2}\|Ax-b\|_2^2+\mu\|x\|_1$
  - $\mu TV(u)+\frac{1}{2}\|u-b\|_F^2$

Funções convexas não diferenciáveis

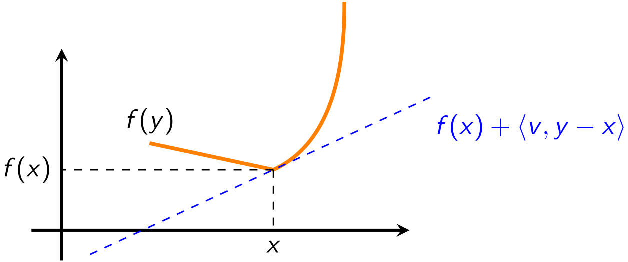
Generalização da noção de derivada: Se $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, $v \in \mathbb{R}^n$ é um **subgradiente** de f no ponto $x \in \mathbb{R}^n$ se

$$f(y) \geq f(x) + \langle v, y - x \rangle, \qquad \forall y \in \mathbb{R}^n.$$

Funções convexas não diferenciáveis

Generalização da noção de derivada: Se $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, $v \in \mathbb{R}^n$ é um **subgradiente** de f no ponto $x \in \mathbb{R}^n$ se

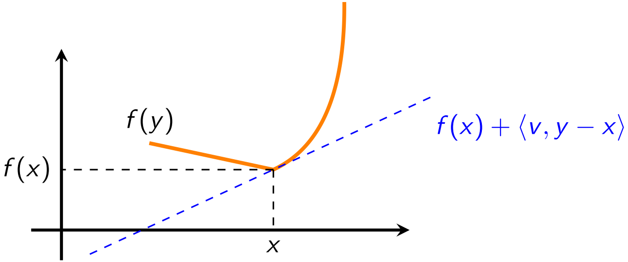
$$f(y) \geq f(x) + \langle v, y - x \rangle, \qquad \forall y \in \mathbb{R}^n.$$



Funções convexas não diferenciáveis

Generalização da noção de derivada: Se $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, $v \in \mathbb{R}^n$ é um **subgradiente** de f no ponto $x \in \mathbb{R}^n$ se

$$f(y) \geq f(x) + \langle v, y - x \rangle, \qquad \forall y \in \mathbb{R}^n.$$

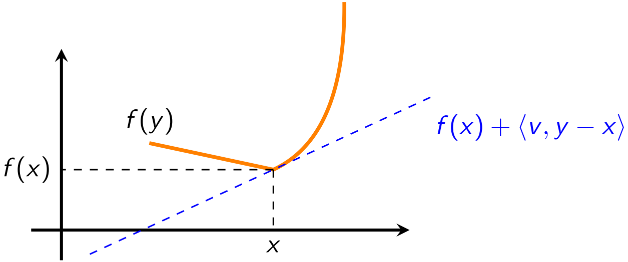


- O subgradiente sempre existe.

Funções convexas não diferenciáveis

Generalização da noção de derivada: Se $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, $v \in \mathbb{R}^n$ é um **subgradiente** de f no ponto $x \in \mathbb{R}^n$ se

$$f(y) \geq f(x) + \langle v, y - x \rangle, \qquad \forall y \in \mathbb{R}^n.$$



- O subgradiente sempre existe.
- *Condição de otimalidade:* $x^\star$ é um mínimo de f se e somente se 0 é um subgradiente de f em $x^\star$.

Método do subgradiente

- $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa não diferenciável.

Método do subgradiente

- $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa não diferenciável.

Gera uma sequência $(x_k)_{k\in\mathbb{N}}$

- Escolher um ponto inicial $x_0 \in \mathbb{R}^n$
- Iterar: $x_{k+1} = x_k - \alpha v_k, \quad k = 0, 1, 2, \ldots,$
  v_k é subgradiente de f em x_k
- $\alpha > 0$ é o comprimento do passo.

Método do subgradiente

- $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa não diferenciável.

Gera uma sequência $(x_k)_{k \in \mathbb{N}}$

- Escolher um ponto inicial $x_0 \in \mathbb{R}^n$
- Iterar: $x_{k+1} = x_k - \alpha v_k, \quad k = 0, 1, 2, \ldots,$
  v_k é subgradiente de f em x_k
- $\alpha > 0$ é o comprimento do passo.

- O método é similar ao método do gradiente, mas substituindo gradientes por subgradientes.

Taxa de convergência

O método do subgradiente não é necessariamente um método de decida, logo guardamos em cada iteração o melhor iterado

$$f(x_k^{best}) = \min_{i=1,\ldots,k} f(x_k)$$

Taxa de convergência

O método do subgradiente não é necessariamente um método de decida, logo guardamos em cada iteração o melhor iterado

$$f(x_k^{best}) = \min_{i=1,\ldots,k} f(x_k)$$

- O método do subgradiente tem taxa de convergência $\mathcal{O}(1/\sqrt{k})$.

  Ou seja, para obter $f(x_k^{best}) - f^\star \leq \epsilon$ é necessário $\mathcal{O}(1/\epsilon^2)$ iterações.

- A taxa de convergência $\mathcal{O}(1/\sqrt{k})$ é muito lenta!

- A taxa de convergência $\mathcal{O}(1/\sqrt{k})$ é muito lenta!

Pode-se melhorar?

- A taxa de convergência $\mathcal{O}(1/\sqrt{k})$ é muito lenta!

Pode-se melhorar? Em geral não! [Nes04]

- A taxa de convergência $\mathcal{O}(1/\sqrt{k})$ é muito lenta!

Pode-se melhorar? Em geral não! [Nes04]

$$\frac{1}{2}\|Ax-b\|_2^2+\mu\|x\|_1$$

- A taxa de convergência $\mathcal{O}(1/\sqrt{k})$ é muito lenta!

Pode-se melhorar? Em geral não! [Nes04]

$$\frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_1$$

Considerando funções da forma

$$f(x) = h(x) + g(x)$$

onde h é convexa e diferenciável e g é convexa, a taxa de convergência $\mathcal{O}(1/k)$ do método do gradiente pode ser recuperada com um algoritmo simples.

Operador proximal

Se $g : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, dado $\alpha > 0$ definimos $\mathbf{prox}_{\alpha g} : \mathbb{R}^n \to \mathbb{R}^n$

$$\mathbf{prox}_{\alpha g}(z) := \arg\min_{x \in \mathbb{R}^n} \alpha g(x) + \frac{1}{2}\|x - z\|^2$$

Operador proximal

Se $g : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, dado $\alpha > 0$ definimos $\mathbf{prox}_{\alpha g} : \mathbb{R}^n \to \mathbb{R}^n$

$$\mathbf{prox}_{\alpha g}(z) := \arg\min_{x \in \mathbb{R}^n} \alpha g(x) + \frac{1}{2}\|x - z\|^2$$

- Se C é um conjunto convexo e fechado e $g = \delta_C$, então

$$\mathbf{prox}_g(z) = P_C(z)$$

Operador proximal

Se $g : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, dado $\alpha > 0$ definimos $\mathbf{prox}_{\alpha g} : \mathbb{R}^n \to \mathbb{R}^n$

$$\mathbf{prox}_{\alpha g}(z) := \arg\min_{x \in \mathbb{R}^n} \alpha g(x) + \frac{1}{2}\|x - z\|^2$$

- Se C é um conjunto convexo e fechado e $g = \delta_C$, então

$$\mathbf{prox}_g(z) = P_C(z)$$

- Se $g(z) = \alpha\|z\|_1$, então

$$\mathbf{prox}_g(z) = \mathbf{shrink}(z, \alpha),$$

onde $\mathbf{shrink}(z, \alpha)_i := \max\{|z_i| - \alpha, 0\}\mathrm{sign}(z_i)$.

Método do gradiente proximal

- $f(x) = h(x) + g(x)$ é tal que h, g são convexas e h é diferenciável com gradiente L-Lipschitz contínuo.

Método do gradiente proximal

- $f(x) = h(x) + g(x)$ é tal que h, g são convexas e h é diferenciável com gradiente L-Lipschitz contínuo.

Gera uma sequência $(x_k)_{k\in\mathbb{N}}$

- Escolher um ponto inicial $x_0 \in \mathbb{R}^n$
- Iterar: $x_{k+1} = \mathbf{prox}_{\alpha g}(x_k - \alpha\nabla h(x_k)), \quad k = 0, 1, 2, \dots,$
- $\alpha > 0$ é o comprimento do passo.

Método do gradiente proximal

- $f(x) = h(x) + g(x)$ é tal que h, g são convexas e h é diferenciável com gradiente L-Lipschitz contínuo.

Gera uma sequência $(x_k)_{k \in \mathbb{N}}$

- Escolher um ponto inicial $x_0 \in \mathbb{R}^n$
- Iterar: $x_{k+1} = \mathbf{prox}_{\alpha g}(x_k - \alpha \nabla h(x_k)), \quad k = 0, 1, 2, \ldots,$
- $\alpha > 0$ é o comprimento do passo.

Taxa de convergência

- Se $\alpha \leq 1/L$, o método do gradiente proximal tem taxa de convergência $\mathcal{O}(1/k)$ ([BeT09], [MoS10]).

É recuperada a taxa de convergência do método do gradiente!

Podemos fazer melhor?

Podemos fazer melhor?

Em [Nes04], foi provado que $\mathcal{O}(1/k^2)$ é uma cota inferior para a taxa de convergência de métodos que têm apenas informação sobre o gradiente em iterações consecutivas:

$$x_k \in x_0 + \mathrm{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\}.$$

Podemos fazer melhor?

Em [Nes04], foi provado que $\mathcal{O}(1/k^2)$ é uma cota inferior para a taxa de convergência de métodos que têm apenas informação sobre o gradiente em iterações consecutivas:

$$x_k \in x_0 + \mathrm{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\}.$$

- Existe método ótimo?

Podemos fazer melhor?

Em [Nes04], foi provado que $\mathcal{O}(1/k^2)$ é uma cota inferior para a taxa de convergência de métodos que têm apenas informação sobre o gradiente em iterações consecutivas:

$$x_k \in x_0 + \mathrm{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\}.$$

- Existe método ótimo?

  Sim! Os métodos do gradiente e do gradiente proximal podem ser acelerados para obter a taxa de convergência $\mathcal{O}(1/k^2)$.

Método do gradiente acelerado

- $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa diferenciável com gradiente L-Lipschitz contínuo.

[Nes83], [Nes04]

- Escolher ponto inicial $x_0 = x_{-1} \in \mathbb{R}^n$ e $t_1 = 1$
- Iterar:
  - $t_{k+1} = \dfrac{1 + \sqrt{4t_k^2 + 1}}{2}$
  - $y_k = x_k + \dfrac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$
  - $x_{k+1} = y_k - \dfrac{1}{L}\nabla f(y_k)$

Método do gradiente acelerado

- $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa diferenciável com gradiente L-Lipschitz contínuo.

[Nes83], [Nes04]

- Escolher ponto inicial $x_0 = x_{-1} \in \mathbb{R}^n$ e $t_1 = 1$
- Iterar:
  - $t_{k+1} = \dfrac{1 + \sqrt{4t_k^2 + 1}}{2}$
  - $y_k = x_k + \dfrac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$
  - $x_{k+1} = y_k - \dfrac{1}{L}\nabla f(y_k)$

- $f(x_k) - f^\star = \mathcal{O}(1/k^2)$

Método do gradiente proximal acelerado

- $f(x) = h(x) + g(x)$ é tal que h, g são convexas e h é diferenciável com gradiente L-Lipschitz contínuo.

[BeT09] (FISTA)

Método do gradiente proximal acelerado

- $f(x) = h(x) + g(x)$ é tal que h, g são convexas e h é diferenciável com gradiente L-Lipschitz contínuo.

[BeT09] (FISTA)

- Escolher ponto inicial $x_0 = x_{-1} \in \mathbb{R}^n$ e $t_1 = 1$
- Iterar:
  - $t_{k+1} = \dfrac{1 + \sqrt{4t_k^2 + 1}}{2}$
  - $y_k = x_k + \dfrac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$
  - $x_{k+1} = \mathbf{prox}_{1/Lg}(y_k - \dfrac{1}{L}\nabla h(y_k))$
- $f(x_k) - f^\star = \mathcal{O}(1/k^2)$

Métodos inerciais

$$\begin{cases} y_k = x_k + \lambda_k(x_k - x_{k-1}) \\ x_{k+1} = \mathbf{prox}_{\alpha g}(y_k - \alpha \nabla h(y_k)) \end{cases}$$

- $\alpha \in (0, 1/L]$ e $\lambda_k \geq 0$

Métodos inerciais

$$\begin{cases} y_k = x_k + \lambda_k(x_k - x_{k-1}) \\ x_{k+1} = \mathbf{prox}_{\alpha g}(y_k - \alpha \nabla h(y_k)) \end{cases}$$

- $\alpha \in (0, 1/L]$ e $\lambda_k \geq 0$
- FISTA (e o método do gradiente acelerado) corresponde a escolha

$$\lambda_k = \frac{t_k - 1}{t_{k+1}} \quad \text{com} \quad t_{k+1} = \frac{1 + \sqrt{4t_k^2 + 1}}{2}$$

Métodos inerciais

$$\begin{cases} y_k = x_k + \lambda_k(x_k - x_{k-1}) \\ x_{k+1} = \mathbf{prox}_{\alpha g}(y_k - \alpha \nabla h(y_k)) \end{cases}$$

- $\alpha \in (0, 1/L]$ e $\lambda_k \geq 0$
- FISTA (e o método do gradiente acelerado) corresponde a escolha

$$\lambda_k = \frac{t_k - 1}{t_{k+1}} \quad \text{com} \quad t_{k+1} = \frac{1 + \sqrt{4t_k^2 + 1}}{2}$$

- Outras escolhas da sequência de parâmetros inerciais $(\lambda_k)_{k\in\mathbb{N}}$ permite melhorar a velocidade de convergência do método e obter propriedades de convergência de $(x_k)_{k\in\mathbb{N}}$.

[AtC18], [AtC16]

- $\lambda_k = 1 - \dfrac{\lambda}{k}$

[AtC18], [AtC16]

- $\lambda_k = 1 - \dfrac{\lambda}{k}$
    - Se $\lambda > 3$, então
      $$f(x_k) - f^\star = o(1/k^2)$$
      e a sequência $(x_k)_{k\in\mathbb{N}}$ converge [CaF11].

[AtC18], [AtC16]

- $\lambda_k = 1 - \dfrac{\lambda}{k}$
  - Se $\lambda > 3$, então
    $$f(x_k) - f^\star = o(1/k^2)$$
    e a sequência $(x_k)_{k\in\mathbb{N}}$ converge [CaF11].
  - Se $\lambda = 3$, a sequência λ_k se comporta similar a sequência de parâmetros inerciais de FISTA e a taxa de convergência é $\mathcal{O}(1/k^2)$.

[AtC18], [AtC16]

- $\lambda_k = 1 - \dfrac{\lambda}{k}$
  - Se $\lambda > 3$, então
    $$f(x_k) - f^\star = o(1/k^2)$$
    e a sequência $(x_k)_{k\in\mathbb{N}}$ converge [CaF11].
  - Se $\lambda = 3$, a sequência λ_k se comporta similar a sequência de parâmetros inerciais de FISTA e a taxa de convergência é $\mathcal{O}(1/k^2)$.
  - A convergência da sequência $(x_k)_{k\in\mathbb{N}}$ gerada por FISTA é um problema em aberto.

"Inexact proximal ϵ-subgradient methods for composite convex optimization problems" (with R.D. Millan)

- Método do gradiente proximal acelerado com cálculo aproximado do operador proximal.

"Inexact proximal ϵ-subgradient methods for composite convex optimization problems" (with R.D. Millan)

- Método do gradiente proximal acelerado com cálculo aproximado do operador proximal.
  - erro relativo ([Bel20])
  - erro absoluto e aproximação do gradiente [Sch11]
  - erro absoluto [Vil13]

"An inertial projective splitting method for the sum of two maximal monotone operators."

- Métodos de decomposição inerciais para resolver problemas de inclusão monótona.

"An inertial projective splitting method for the sum of two maximal monotone operators."

- Métodos de decomposição inerciais para resolver problemas de inclusão monótona.
  - cálculo aproximado dos subproblemas proximais com erro relativo
  - convergência da sequência gerada pelo método
  - taxa de convergência para soluções aproximadas do problema de inclusão monótona
  - quando aplicado a otimização convexa, qual é a taxa de convergência dos valores funcionais?

Obrigada!

Referências:

[AtC18] H. Attouch, and A. Cabot. Convergence rates of inertial Forward-backward algorithms. Siam J. Optim., 28 (2018), 849–874.

[AtP16] H. Attouch and J. Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $\frac{1}{k^2}$. SIAM J. Optim., 26 (2016), 1824–1834.

[Bel20] Bello Cruz, Y.; Gonçalves, M.L.N.; Krislock, N.: On inexact accelerated proximal gradient methods with relative error rules, (2020)

[CaF11] A. Cabot and P. Frankel, Asymptotics for some proximal-like method involving inertia and memory aspects, Set Valued Var. Anal., 19 (2011), 59–74.

Referências:

[MiM19] R.D. Millán and M.P. Machado. Inexact proximal ϵ-subgradient methods for composite convex optimization problems. J Glob Optim (2019). https://doi.org/10.1007/s10898-019-00808-8

[Mac23] M. P. Machado. An inertial projective splitting method for the sum of two maximal monotone operators (2023).

[Nes83] Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Sov.Math. Dokl.27, 372–376 (1983)

[Nes04] Nesterov, Y.: Introductory Lectures on Convex Optimization. Kluwer, Boston, 2004

Referências:

[Sch11] Schmidt, M., Le Roux, N., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: NIPS'11—25th Annual Conference on Neural Information Processing Systems(Grenada, Spain, Dec. 2011)

[Vil13] S. Villa, S. Salzo, L. Baldassarre and A. Verri. Accelerated and inexact forward-backward algorithms. JSIAM J. Optim., 23(3) (2013), 1607–1633.