Improving “Fast Iterative Shrinkage-Thresholding Algorithm”

Faster, Smarter and Greedier

Jingwei Liang

Department of Applied Mathematics and Theoretical Physics

Joint work with: Carola Schönlieb (Cambridge)

Table of contents

Optimisation problem

Composite optimisation problem

$$\min_{x\in\mathbb{R}^n} \left\{\Phi(x) = R(x) + F(x)\right\}.$$

Assumptions

- R is proper convex and lower semi-continuous.
- F is convex, differentiable, and ∇F is L-Lip. continuous.
- $\mathrm{Argmin}(\Phi) \neq \emptyset$, *i.e.* the set of minimisers is non-empty.

Forward–Backward splitting [Lions & Mercier, 1979]

Let $\gamma_k \in [0, 2/L]$:

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla F(x_k)\big),$$

where

$$\mathrm{prox}_{\gamma R}(\cdot) \stackrel{\mathrm{def}}{=} \mathrm{argmin}_{x \in \mathbb{R}^n}\, \gamma R(x) + \tfrac{1}{2}\|x - \cdot\|^2.$$

Special cases Gradient descent $R = 0$, proximal point algorithm $F = 0$.

Convergence $0 < \underline{\gamma} \leq \gamma_k \leq \bar{\gamma} < 2/L$ [Combettes & Wajs, 2005]

Con. rate $\Phi(x_k) - \Phi(x^\star) = o(1/k)$ [Molinari et al, 18].

$\|x_k - x_{k-1}\| = o(1/\sqrt{k})$ [Liang et al, 16].

Liner convergence under: PL, QG, SC...

FISTA-BT [Beck & Teboulle, 2009]

Let $\gamma_k \in]0, 1/L]$ and $t_0 = 1$:

$$t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}, \quad a_k = \frac{t_{k-1}-1}{t_k},$$
$$y_k = x_k + a_k(x_k - x_{k-1}),$$
$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(y_k - \gamma_k \nabla F(y_k)\big).$$

Con. rate $\Phi(x_k) - \Phi(x^\star) = O(1/k^2)$ (optimal).

Open problem: $x_k \to x^\star$?

FISTA-CD $t_k = \frac{k+d-1}{d}, a_k = \frac{t_{k-1}-1}{t_k}, d > 2$ [Chambolle & Dossal, 2015].

$\Phi(x_k) - \Phi(x^\star) = o(1/k^2)$ [Attouch & Peypouquet, 16].

$\|x_k - x_{k-1}\| = o(1/k)$ [Attouch & Peypouquet, 16].

Problems

- Convergence of $\{x_k\}_{k\in\mathbb{N}}$ for FISTA-BT remains unclear.
- For FISTA-CD, *i.e.* $t_k = \frac{k+d-1}{d}$, $a_k = \frac{t_{k-1}-1}{t_k}$. For

$$d = 2 \quad \text{and} \quad d = 20,$$

  significantly different performances. Why?
- Strong convexity $\rightarrow$ optimal inertial parameters. However,
  - very often strong convexity is not available and time consuming to estimate?
  - does optimal scheme really mean the fastest in practice?
- Oscillation $\rightarrow$ Restarting FISTA, can we further improve it?

Outline

Recall that t_k of FISTA-BT

$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2}.$$

Observation I

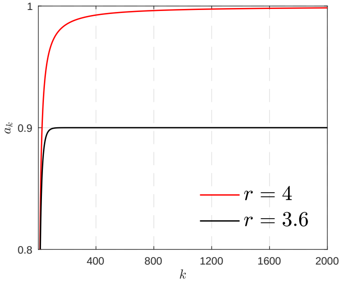
Replace 4 with $r > 0$,

$$t_k = \frac{1 + \sqrt{1 + rt_{k-1}^2}}{2}.$$

Then

$$r \in]0, 4[\,:\, t_k \to \bar{t} < +\infty,\ a_k \to \bar{a} < 1,$$

$$r = 4 \,:\, t_k \approx \frac{k+1}{2},\ a_k \to 1.$$



Now

$$t_k = \frac{1 + \sqrt{1 + r t_{k-1}^2}}{2}.$$

Observation II

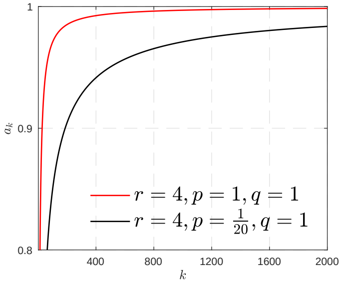
Replace 1 with $p, q > 0$,

$$t_k = \frac{p + \sqrt{q + r t_{k-1}^2}}{2}.$$

Then,

$$r \in]0, 4[: t_k \to \bar{t} < +\infty, \ \ a_k \to \bar{a} < 1,$$

$$r = 4 : t_k \approx \frac{k+1}{2} p, \ \ a_k \to 1.$$



Algorithm - FISTA-Mod

Let $\gamma_k \in]0, 1/L]$, $p, q \in]0, 1]$, $r \in]0, 4]$ and $t_0 = 1$:

$$t_k = \frac{p + \sqrt{q + r t_{k-1}^2}}{2}, \quad a_k = \frac{t_{k-1} - 1}{t_k}$$

$$y_k = x_k + a_k(x_k - x_{k-1})$$

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(y_k - \gamma_k \nabla F(y_k)\big)$$

FISTA-BT $(p, q, r) = (1, 1, 4)$.

Inertial FB $r < 4$.

Objective function

For FISTA-Mod, let $r = 4$ and $p \in]0, 1]$, $q \leq (2-p)^2$, then

$$\Phi(x_k) - \Phi(x^\star) \leq \frac{2L}{p^2(k+1)^2}\|x_0 - x^\star\|^2.$$

If moreover $p < 1$ and $q \in [p^2, (2-p)^2]$, $\Phi(x_k) - \Phi(x^\star) = o(1/k^2)$.

FISTA-BT: $t_k^2 - t_k = t_{k-1}^2$,

- For $q \leq (2-p)^2$: $t_k^2 - t_k \leq t_{k-1}^2$.
- For $p < 1$ and $q \in [p^2, (2-p)^2]$:

$$\frac{p(1-p)(k+1)}{2} \leq t_{k-1}^2 - (t_k^2 - t_k).$$

Objective function

For FISTA-Mod, let $r = 4$ and $p \in]0, 1]$, $q \leq (2 - p)^2$, then

$$\Phi(x_k) - \Phi(x^\star) \leq \frac{2L}{p^2(k+1)^2}\|x_0 - x^\star\|^2.$$

If moreover $p < 1$ and $q \in [p^2, (2-p)^2]$, $\Phi(x_k) - \Phi(x^\star) = o(1/k^2)$.

Sequence

For FISTA-Mod, let $r = 4$ and $p \in]0, 1[$, $q \in [p^2, (2-p)^2]$. Then

- $x_k \to x^\star \in \mathrm{Argmin}(\Phi)$.
- $\|x_k - x_{k-1}\| = o(1/k)$.

Outline

Lazy-start FISTA

FISTA-Mod $p \in [\frac{1}{80}, \frac{1}{10}]$, $q \in [0, 1]$ and $r = 4$;
FISTA-CD $d \in [10, 80]$.

Consider

$$\min_{x \in \mathbb{R}^{201}} \frac{1}{2}\|Ax - y\|^2,$$

where $y = 0$ and

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \cdots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}_{201 \times 201}.$$

We have

$$L = 16, \alpha = 5.8 \times 10^{-8}.$$

Lazy-start FISTA

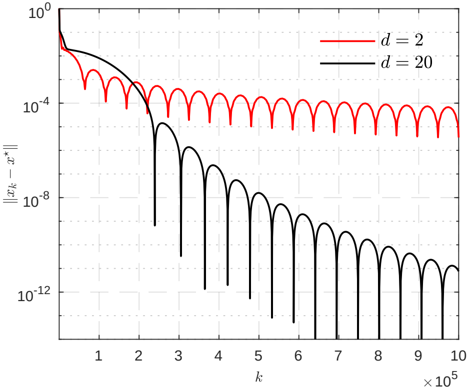
FISTA-Mod $p \in [\frac{1}{80}, \frac{1}{10}], q \in [0, 1]$ and $r = 4$;

FISTA-CD $d \in [10, 80]$.

FISTA-CD: $\gamma = 1/L$, and

$$t_k = \frac{k+d-1}{d}, \quad a_k = \frac{k-1}{k+d},$$

with $d = 2$ and $d = 20$.



“A small leak will sink a great ship.”

– Benjamin Franklin

Some notations

- Lipschitz constant L.
- Strong convexity α.
- Step-size $\gamma = 1/L$.
- Fixed-point matrix of gradient descent: $G = \mathrm{Id} - \gamma A^T A$.
- Leading eigenvalue of G: $\eta = 1 - \gamma\alpha$.
- Optimal inertial parameter: $a^\star = \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}}$.
- Optimal convergence rate: $\rho^\star = 1 - \sqrt{\gamma\alpha}$.

Fixed-point characterisation

- Since $y_{k-1} = x_{k-1} + a_{k-1}(x_{k-1} - x_{k-2})$,

$$\begin{aligned} x_k - x^\star &= G(y_{k-1} - x^\star) \\ &= (1 + a_{k-1})G(x_{k-1} - x^\star) - G(x_{k-2} - x^\star). \end{aligned}$$

- Define

$$z_k \stackrel{\text{def}}{=} \begin{pmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{pmatrix} \quad \text{and} \quad M_{k-1} \stackrel{\text{def}}{=} \begin{bmatrix} (1 + a_{k-1})G & -a_{k-1}G \\ \mathrm{Id} & 0 \end{bmatrix}.$$

- Then

$$z_k = M_{k-1} z_{k-1}.$$

- Let $\widetilde{M}_k \stackrel{\text{def}}{=} \prod_{i=1}^{k-1} M_{k-i}$,

$$z_k = \widetilde{M}_k z_1.$$

Spectral property of $\widetilde{M}_k$

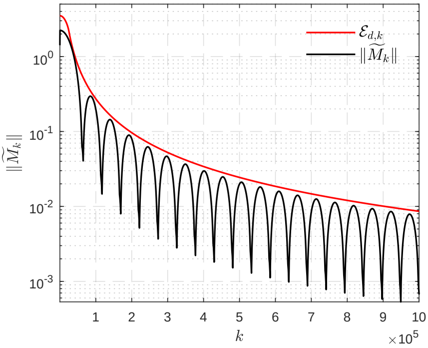
NB: $a^\star = \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}}$ is the optimal inertial parameter

- Denote σ_k the leading eigenvalue of M_k,
$$\sigma_k \in \begin{cases} \mathbb{R} : a_k \le a^\star, \\ \mathbb{C} : a_k \ge a^\star. \end{cases}$$
Let $\rho_k = |\sigma_k|$.

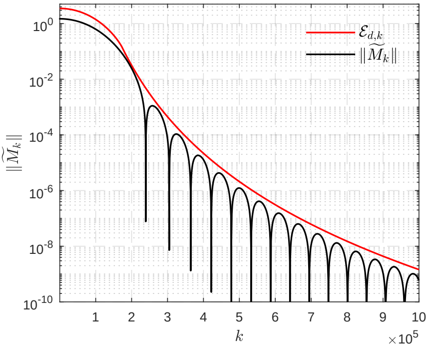
- There exists $\mathcal{T} > 0$, such that
$$\|\widetilde{M}_k\| \le \mathcal{T} \prod_{i=1}^{k-1} \rho_{k-i}.$$

- Let $\widetilde{\mathcal{T}}$ be the minimal value such that the above holds, denote
$$\mathcal{E}_{d,k} \stackrel{\text{def}}{=} \widetilde{\mathcal{T}} \prod_{i=1}^{k-1} \rho_{k-i}$$
the **envelope** of $\|\widetilde{M}_k\|$.

Spectral property of $\widetilde{M}_k$

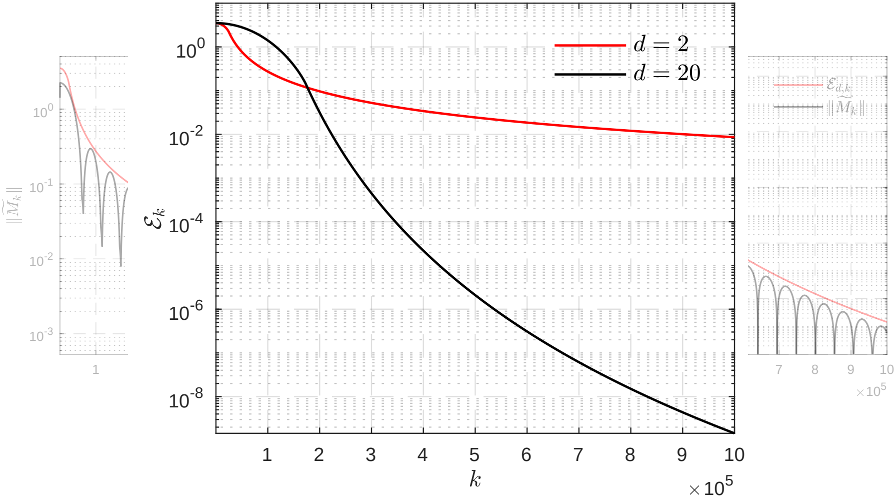


$d = 2$

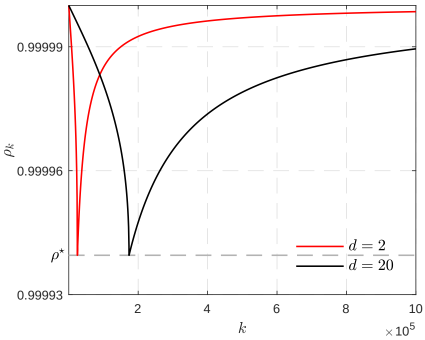


$d = 20$

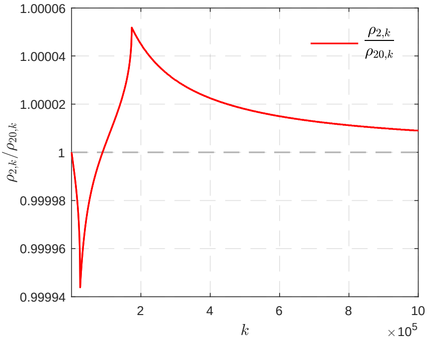
NB: $\widetilde{\mathcal{T}}$ is the same for the two cases.



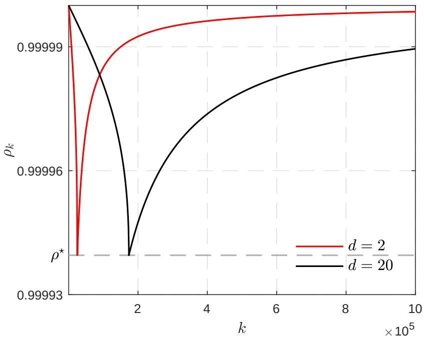
Comparison of $\mathcal{E}_{d,k}$



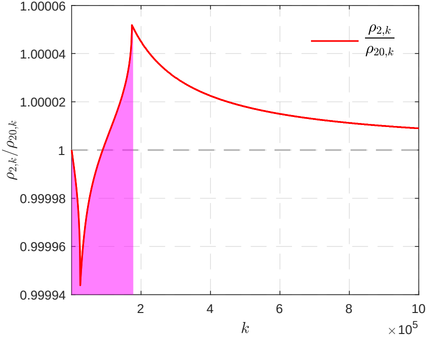
(a) ρ_k



(b) $\rho_{2,k}/\rho_{20,k}$



(a) ρ_k

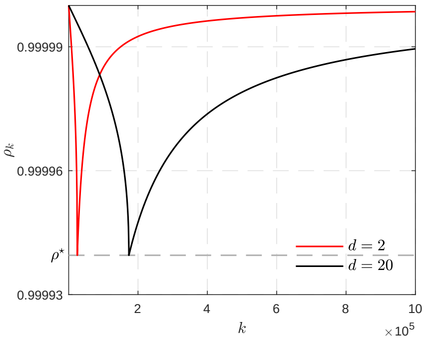


(c) Integral to K_{eq}

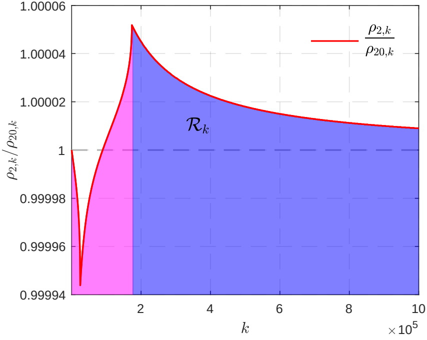
$\mathcal{E}_{2,k} = \mathcal{E}_{20,k}$ for

$$k \approx K_{\text{eq}} \stackrel{\text{def}}{=} \left\lceil \frac{1 + 20a^\star}{1 - a^\star} \right\rceil + 1.$$

The small leak



(a) ρ_k



(d) $\mathcal{R}_k$

For $k \geq K_{\text{eq}}$,

$$\frac{|\rho_{2,k}|}{|\rho_{20,k}|} = \sqrt{\frac{k+20}{k+2}}.$$

Let $d_1 = 2, d_2 = 20$ and $k \geq K_{\mathrm{eq}} + 2(d_2 - d_1)$, then

$$
\begin{aligned}
\mathcal{R}_k &= \prod_{i=K_{\mathrm{eq}}}^{k} \frac{\sqrt{a_{d_1,i}}}{\sqrt{a_{d_2,i}}} \\
&= \prod_{i=K_{\mathrm{eq}}}^{k} \sqrt{\frac{i+d_2}{i+d_1}} \\
&= \prod_{j=0}^{d_2-d_1-1} \left(\frac{k+d_1+1+j}{K_{\mathrm{eq}}+d_1+j}\right)^{1/2} \\
&\approx \left(\frac{k+d_2}{K_{\mathrm{eq}}+d_2-1}\right)^{(d_2-d_1)/2} .
\end{aligned}
$$

Now let's sink the ship...

Denote $\mathcal{C} \stackrel{\text{def}}{=} L/\alpha$,

$$\mathcal{R}_k \approx \left(\frac{2}{\sqrt{\mathcal{C}}+1}\right)^{(d_2-d_1)/2} \left(\frac{k+d_2}{1+d_2}\right)^{(d_2-d_1)/2}.$$

- For $n = 201$,

$$L = 16 \quad \text{and} \quad \alpha = 5.85 \times 10^{-8},$$

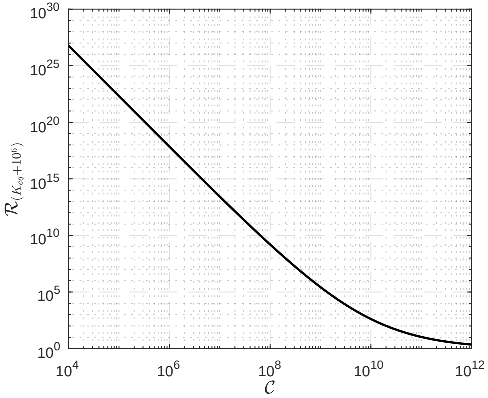
  and $\mathcal{C} = 2.735 \times 10^8$.

- Let $k = 10^6$, we have $\mathcal{R}_k \approx 5.98 \times 10^6$, while for $\mathcal{E}_{d,k}$

$$\frac{\mathcal{E}_{d_1,k=10^6}}{\mathcal{E}_{d_2,k=10^6}} = 5.96 \times 10^6.$$

Given $\mathcal{C} \in [10^4, 10^{12}]$, value of

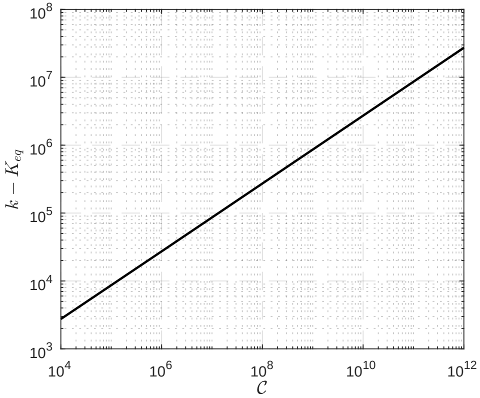
$$\mathcal{R}_{K_{\mathrm{eq}}+10^6}.$$



Given $\mathcal{C} \in [10^4, 10^{12}]$, value of

$$k - K_{\mathrm{eq}}$$

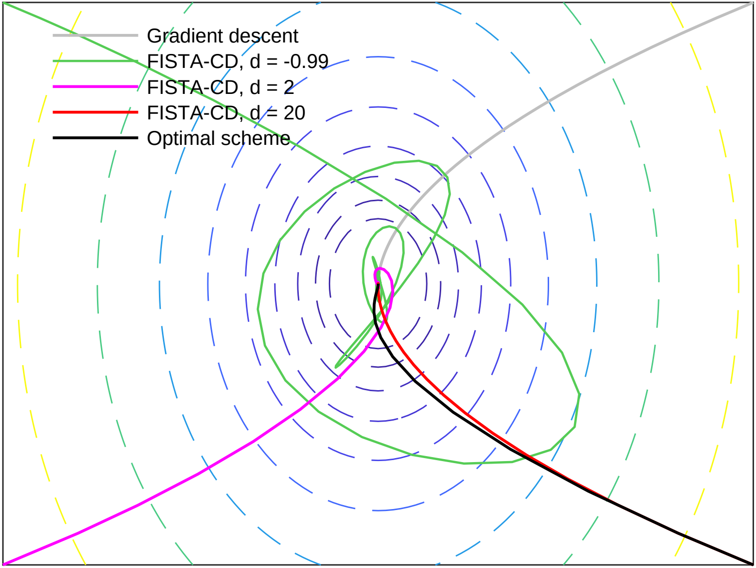
such that $\mathcal{R}_k = 10^5$...



Remarks

- There exists optimal choice for d.
- The optimal choice of d is independent of condition number $\mathcal{C}$.
- But depends on stopping tolerence.
- The discussion is via $\|x_k - x^\star\|$, and in practice only $\|x_k - x_{k-1}\|$ available...

Trajectory of gradient descent and FISTA



Adaptive to strong convexity

Assume that

$$F \text{ is } \alpha\text{-strongly convex and } R \text{ is only convex.}$$

Given γ, the optimal inertial parameter

$$a^\star = \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}}.$$

For FISTA-Mod, with $p, q \in [0, 1]$, the optimal value of r reads

$$\begin{aligned} r = f(\alpha, \gamma; p, q) &= 4(1-p) + 4pa^\star + (p^2 - q)(1-a^\star)^2 \\ &= 4(1-p) + \frac{4p(1-\sqrt{\gamma\alpha})}{1+\sqrt{\gamma\alpha}} + \frac{4\gamma\alpha(p^2-q)}{(1+\sqrt{\gamma\alpha})^2} \leq 4. \end{aligned}$$

Algorithm - α-FISTA

Let $p, q > 0$ and $\gamma \leq 1/L$. For $\alpha \geq 0$,

$$r = 4(1-p) + \frac{4p(1-\sqrt{\gamma\alpha})}{1+\sqrt{\gamma\alpha}} + \frac{4\gamma\alpha(p^2-q)}{(1+\sqrt{\gamma\alpha})^2}.$$

Let $t_0 \geq 1$, and $x_0 \in \mathbb{R}^n, x_{-1} = x_0$:

$$\left\lfloor \begin{aligned} t_k &= \frac{p+\sqrt{q+rt_{k-1}^2}}{2}, \quad a_k = \frac{t_{k-1}-1}{t_k}, \\ y_k &= x_k + a_k(x_k - x_{k-1}), \\ x_{k+1} &= \mathrm{prox}_{\gamma R}\big(y_k - \gamma\nabla F(y_k)\big). \end{aligned} \right.$$

- There holds

$$\Phi(x_k) - \Phi(x^\star) \leq \mathcal{C}\omega_k,$$

  where $\mathcal{C} > 0$ is a constant and $\omega_k = \min\left\{\frac{2L}{p^2(k+1)^2}, (1-\sqrt{\gamma\alpha})^k\right\}$;

- In practice, better to choose $t_0 > \dfrac{2p+\sqrt{rp^2+(4-r)q}}{4-r}$.

However,

- ☹ In practice, only locally strongly convex, or simply convex;
- ☹ Oscillation is independent of strong convexity...

Algorithm - Rada-FISTA

Let $\gamma \in]0, 1/L]$, $\xi < 1$, $p, q \in]0, 1]$, $r = 4$ and $t_0 = 1$:

- Run FISTA-Mod:

$$t_k = \frac{p + \sqrt{q + rt_{k-1}^2}}{2}, \quad a_k = \frac{t_{k-1} - 1}{t_k},$$
$$y_k = x_k + a_k(x_k - x_{k-1}),$$
$$x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma \nabla F(y_k)\big).$$

- If $(y_k - x_{k+1})^T(x_{k+1} - x_k) > 0$:
  - Option I: $r = \xi r$, $y_k = x_k$;
  - Option II: $r = \xi r$, $y_k = x_k$ and $t_k = 1$.

It is a circle:

$$a_k \nearrow 1 \rightarrow \text{close to } 1 \rightarrow \text{oscillation} \rightarrow \text{restarting} \rightarrow a_k \nearrow 1 \ldots$$

Very often, resetting a_k to 0 is not a good idea...

Algorithm - Greedy FISTA

Let $\gamma \in [1/L, 2/L[$, $\xi < 1, S > 1, p, q \in]0, 1], r = 4$ and $t_0 = 1$:

- Run iteration:

$$y_k = x_k + (x_k - x_{k-1}),$$
$$x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma \nabla F(y_k)\big).$$

- Restarting: if $(y_k - x_{k+1})^T(x_{k+1} - x_k) > 0$, $y_k = x_k$;
- Safeguard: if $\|x_{k+1} - x_k\| \geq S\|x_1 - x_0\|$, $\gamma = \max\{\xi\gamma, \frac{1}{L}\}$;

Outline

Methods for comparisons

- The original FISTA-BT [Beck & Teboulle, 2009].
- FISTA-Mod with $p = 1/20$ and $q = 1/2$.
- The original restarting FISTA [O'Donoghue, 2012].
- Rada-FISTA.
- Greedy FISTA with $\gamma = 1.3/L$ and $S = 1, \xi = 0.96$.

Regularised least square

$$\min_{x\in\mathbb{R}^n} R(x) + \frac{1}{2}\|Kx - f\|^2.$$

Two different R:

$$\begin{aligned} \ell_\infty\text{-norm} \quad & (m,n) = (1020, 1024),\ x_{\mathrm{ob}} \text{ has 32 saturated entries.} \\ \text{Total variation} \quad & (m,n) = (256, 1024),\ \nabla x_{\mathrm{ob}} \text{ is 32-sparse.} \end{aligned}$$

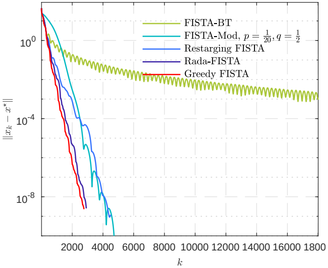
Sparse logistic regression

$$\min_{x\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{m}\sum_{i=1}^{m} \log\big(1 + e^{-l_i h_i^T x}\big),$$

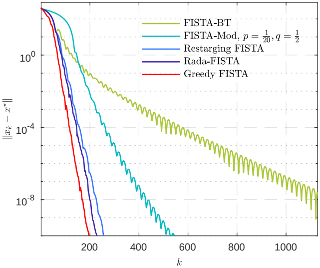
where $\mu = 10^{-2}$. The `australian` data set from LIBSVM[1] is considered.

[1] `https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/`

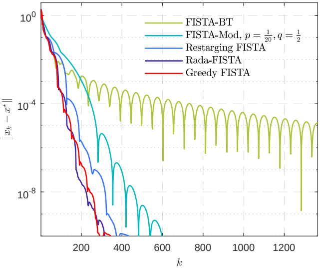
Regression problems



ℓ_∞-norm

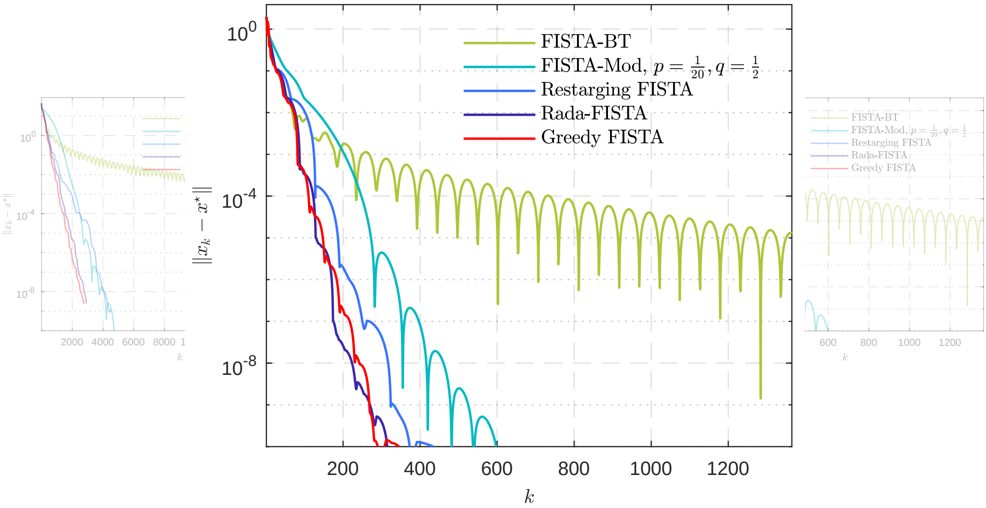


Total variation



SLR

Regression problems



Sparse LR

Principal component pursuit (PCP) [Candès et al, 2011]

$$\min_{x_l, x_s \in \mathbb{R}^{m\times n}} \lambda_1 \|x_s\|_1 + \lambda_2 \|x_l\|_* + \frac{1}{2}\|f - x_l - x_s\|_F^2.$$

For fixed x_l, we have $x_s^\star = \mathrm{prox}_{\mu\|\cdot\|_1}(f - x_l)$. Thus, PCP is equivalent to

$$\min_{x_l \in \mathbb{R}^{m\times n}} {}^{1}\big(\mu\|\cdot\|_1\big)(f - x_l) + \nu\|x_l\|_*,$$

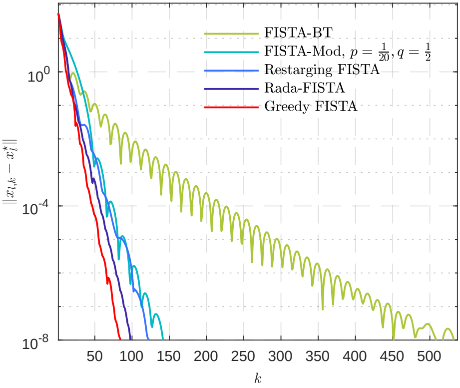
where

$${}^{1}\big(\mu\|\cdot\|_1\big)(f - x_l) = \min_z \frac{1}{2}\|f - x_l - z\|_F^2 + \mu\|z\|_1$$

is the Moreau Envelope of $\mu\|\cdot\|_1$ of index 1, and has 1-Lipschitz continuous gradient.

Principal component pursuit (PCP)

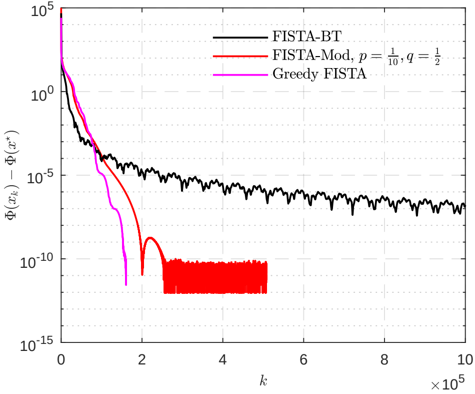
$$\min_{x_l \in \mathbb{R}^{m\times n}} {}^{1}\big(\mu\|\cdot\|_1\big)(f - x_l) + \nu\|x_l\|_* .$$



Non-negative least square

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Kx - f\|^2 \quad \text{s.t.} \quad x_i \geq 0.$$

Challenge: $\mathtt{cond}(K) = O(10^{18})$.



Outline

Takeaway messages

A modified FISTA scheme with sequence convergence guarantee

Lazy-start, adaptive, restarting acceleration

Superior performance over original FISTA

Takeaway messages

A modified FISTA scheme with sequence convergence guarantee

Lazy-start, adaptive, restarting acceleration

Superior performance over original FISTA

Thank you very much!

https://github.com/jliang993/Faster-FISTA