Adaptive FISTA

Peter Ochs
Saarland University

— 07.06.2018 —

joint work with Thomas Pock, TU Graz, Austria

Some Facts about FISTA

Some Facts about FISTA:

- FISTA developed in [Beck, Teboulle: *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM journal on imaging sciences 2(1):183–202, 2009], **5600 citations on google scholar**.

Some Facts about FISTA:

- FISTA developed in [Beck, Teboulle: *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM journal on imaging sciences 2(1):183–202, 2009], **5600 citations on google scholar**.
- Motivated by [Nesterov: *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Soviet Mathematics Doklady, 1983], **2000 citations on google scholar**.

⇝ "**optimal method**" in the sense of [Nemirovskii, Yudin '83], [Nesterov '04].

Some Facts about FISTA

Some Facts about FISTA:

- FISTA developed in [Beck, Teboulle: *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM journal on imaging sciences 2(1):183–202, 2009], **5600 citations on google scholar**.

- Motivated by [Nesterov: *A method of solving a convex programming problem with convergence rate $O(1/k^2)$*, Soviet Mathematics Doklady, 1983], **2000 citations on google scholar**.

⇝ "**optimal method**" in the sense of [Nemirovskii, Yudin '83], [Nesterov '04].

- **Accelerated Gradient Method:**

$$y^{(k)} = x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)}) \qquad \Big((\beta_k)_{k\in\mathbb{N}} \text{ cleverly predefined}\Big)$$
$$x^{(k+1)} = y^{(k)} - \tau\nabla f(y^{(k)})$$

- FISTA extends Accelerated Gradient Method by [Nesterov '83] to non-smooth problem.

A Class of Structured Non-smooth Optimization Problems

A Class of Structured Non-smooth Optimization Problems:

$$\min_x \quad g(x) \quad + \quad f(x)$$

simple proximal mapping — $g(x)$

$\operatorname{argmin}_x g(x) + \frac{1}{2}\|x - \bar{x}\|^2$

smooth — $f(x)$

Update Scheme: FISTA (f, g convex)

$$y_{\beta_k}^{(k)} = x^{(k)} + \beta_k (x^{(k)} - x^{(k-1)})$$

$$x^{(k+1)} = \operatorname*{argmin}_x g(x) + f(y_{\beta_k}^{(k)}) + \left\langle \nabla f(y_{\beta_k}^{(k)}), x - y_{\beta_k}^{(k)} \right\rangle + \frac{1}{2\tau}\|x - y_{\beta_k}^{(k)}\|^2$$

A Class of Structured Non-smooth Optimization Problems

A Class of Structured Non-smooth Optimization Problems:

$$\min_x \quad g(x) \quad + \quad f(x)$$

simple proximal mapping $g(x)$: $\operatorname{argmin}_x g(x) + \frac{1}{2}\|x - \bar{x}\|^2$

smooth: $f(x)$

Update Scheme: FISTA (f, g convex)

$$y^{(k)}_{\beta_k} = x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)})$$

$$x^{(k+1)} = \operatorname*{argmin}_x g(x) + f(y^{(k)}_{\beta_k}) + \left\langle \nabla f(y^{(k)}_{\beta_k}), x - y^{(k)}_{\beta_k} \right\rangle + \frac{1}{2\tau}\|x - y^{(k)}_{\beta_k}\|^2$$

Equivalent to

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^N} g(x) + \frac{1}{2\tau}\|x - \left(y^{(k)}_{\beta_k} - \tau \nabla f(y^{(k)}_{\beta_k})\right)\|^2 =: \operatorname{prox}_{\tau g}\left(y^{(k)}_{\beta_k} - \tau \nabla f(y^{(k)}_{\beta_k})\right)$$

A Class of Structured Non-smooth Optimization Problems

A Class of Structured Non-smooth Optimization Problems:

$$\min_x \quad g(x) \quad + \quad f(x)$$

simple proximal mapping — $g(x)$

$\operatorname{argmin}_x g(x) + \frac{1}{2}\|x - \bar{x}\|^2$

smooth — $f(x)$

Update Scheme: Adaptive FISTA (also non-convex)

$$y^{(k)}_{\beta_k} = x^{(k)} + \beta_k (x^{(k)} - x^{(k-1)})$$

$$x^{(k+1)} = \operatorname*{argmin}_x \min_{\beta_k} g(x) + f(y^{(k)}_{\beta_k}) + \left\langle \nabla f(y^{(k)}_{\beta_k}), x - y^{(k)}_{\beta_k} \right\rangle + \frac{1}{2\tau}\|x - y^{(k)}_{\beta_k}\|^2$$

A Class of Structured Non-smooth Optimization Problems

A Class of Structured Non-smooth Optimization Problems:

$$\min_x \quad g(x) \quad + \quad f(x)$$

simple proximal mapping (g) — smooth (f)

$$\operatorname{argmin}_x g(x) + \tfrac{1}{2}\|x - \bar{x}\|^2$$

Update Scheme: Adaptive FISTA (f quadratic)

$$y^{(k)}_{\beta_k} = x^{(k)} + \beta_k (x^{(k)} - x^{(k-1)})$$

$$x^{(k+1)} = \operatorname*{argmin}_x \min_{\beta_k} g(x) + f(y^{(k)}_{\beta_k}) + \left\langle \nabla f(y^{(k)}_{\beta_k}), x - y^{(k)}_{\beta_k} \right\rangle + \frac{1}{2\tau}\|x - y^{(k)}_{\beta_k}\|^2$$

... Taylor expansion around $x^{(k)}$ and optimize for $\beta_k = \beta_k(x)$ ...

$$x^{(k+1)} = \operatorname*{argmin}_x \; g(x) + \frac{1}{2}\|x - (x^{(k)} - \boldsymbol{Q}_k^{-1}\nabla f(x^{(k)}))\|^2_{\boldsymbol{Q}_k}$$

A Class of Structured Non-smooth Optimization Problems

Update Scheme: Adaptive FISTA (f **quadratic**)

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^N} g(x) + \frac{1}{2}\|x - (x^{(k)} - \boldsymbol{Q}_k^{-1}\nabla f(x^{(k)}))\|^2_{\boldsymbol{Q}_k}$$
$$=: \operatorname{prox}_g^{\boldsymbol{Q}_k}(x^{(k)} - \boldsymbol{Q}_k^{-1}\nabla f(x^{(k)}))$$

with $\boldsymbol{Q}_k \in \mathbb{S}_{++}(N)$ as in the **(zero memory) SR1 quasi-Newton method**:

$$\boldsymbol{Q} = \boldsymbol{I} - uu^\top \quad \text{(identity minus rank-1)}\,.$$

- SR1 proximal quasi-Newton method proposed by [Becker, Fadili '12] (convex case).
- Special setting is treated in [Karimi, Vavasis '17].
- Unified and extended in [Becker, Fadili, O. '18].

Discussion about Solving the Proximal Mapping

Discussion about Solving the Proximal Mapping: (g convex)

- For general $\boldsymbol{Q}$, the main algorithmic step is hard to solve:

$$\hat{x} = \mathrm{prox}_g^{\boldsymbol{Q}} := \operatorname*{argmin}_{x \in \mathbb{R}^N} g(x) + \frac{1}{2}\|x - \bar{x}\|_{\boldsymbol{Q}}^2$$

- **Theorem:** [Becker, Fadili '12]

  $\boldsymbol{Q} = \boldsymbol{D} \pm uu^\top \in \mathbb{S}_{++}$ for $u \in \mathbb{R}^N$ and $\boldsymbol{D}$ diagonal. Then

$$\mathrm{prox}_g^{\boldsymbol{Q}}(\bar{x}) = \boldsymbol{D}^{-1/2} \circ \mathrm{prox}_{g \circ \boldsymbol{D}^{-1/2}}(\boldsymbol{D}^{1/2}\bar{x} \mp v^\star)$$

  where $v^\star = \alpha^\star \boldsymbol{D}^{-1/2} u$ and $\alpha^\star$ is the unique root of

$$l(\alpha) = \left\langle u, \bar{x} - \boldsymbol{D}^{-1/2} \circ \mathrm{prox}_{g \circ \boldsymbol{D}^{-1/2}} \circ \boldsymbol{D}^{1/2}(\bar{x} \mp \alpha \boldsymbol{D}^{-1} u) \right\rangle + \alpha\,,$$

  which is strictly increasing and Lipschitz continuous with $1 + \sum_i u_i^2 d_i$.

Solving the rank-1 Proximal Mapping for ℓ_1-norm

Example: (Solving the rank-1 prox of the ℓ_1-norm)

- The **proximal mapping wrt. the diagonal matrix** is separable and simple

$$
\begin{aligned}
\operatorname{prox}_{g\circ \boldsymbol{D}^{-1/2}}(z) &= \operatorname*{argmin}_{x\in\mathbb{R}^N} \|\boldsymbol{D}^{-1/2}x\|_1 + \frac{1}{2}\|x-z\|^2 \\
&= \operatorname*{argmin}_{x\in\mathbb{R}^N} \sum_{i=1}^{N} |x_i|/\sqrt{d_i} + \frac{1}{2}(x_i - z_i)^2 \\
&= \Big(\operatorname*{argmin}_{x_i\in\mathbb{R}} |x_i|/\sqrt{d_i} + \frac{1}{2}(x_i - z_i)^2\Big)_{i=1,\dots,N} \\
&= \Big(\max\big(0, |z_i| - 1/\sqrt{d_i}\big)\operatorname{sign}(z_i)\Big)_{i=1,\dots,N}
\end{aligned}
$$

Solving the rank-1 Proximal Mapping for ℓ_1-norm

The root finding problem in the rank-1 prox of the ℓ_1-norm:

- $\alpha^\star$ is the root of the **1D function** (i.e. $l(\alpha^\star) = 0$)

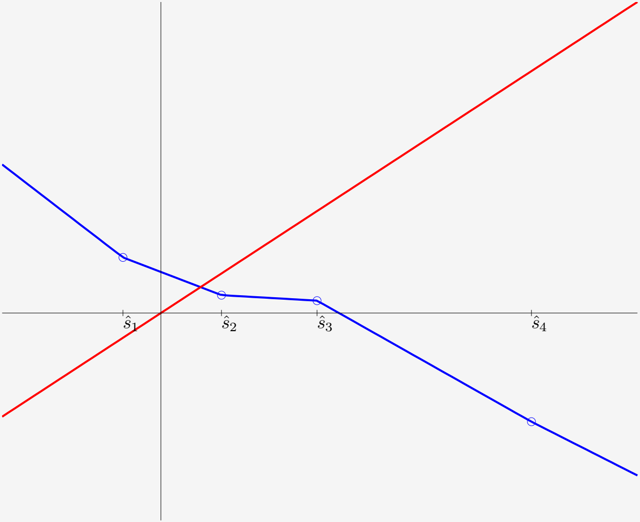
$$
\begin{aligned}
l(\alpha) &= \Big\langle u, \bar{x} - \boldsymbol{D}^{-1/2} \circ \mathrm{prox}_{g \circ \boldsymbol{D}^{-1/2}} \circ \boldsymbol{D}^{1/2}(\bar{x} \mp \alpha \boldsymbol{D}^{-1} u) \Big\rangle + \alpha \\
&= \big\langle u, \bar{x} - \mathrm{PLin}(\bar{x} \mp \alpha \boldsymbol{D}^{-1} u) \big\rangle + \alpha
\end{aligned}
$$

  which is a **piecewise linear function**.

- Construct this function by sorting $K \geq N$ **breakpoints**. Cost: $\mathcal{O}(K \log(K))$.
- The root is determined using **binary search**. Cost: $\mathcal{O}(\log(K))$.
  (*remember:* $l(\alpha)$ is **strictly increasing**)
- Computing $l(\alpha)$ costs $\mathcal{O}(N)$.

$\rightsquigarrow$ **Total cost:** $\mathcal{O}(K \log(K))$.

Solving the rank-1 Proximal Mapping for ℓ_1-norm



from [S. Becker]

Discussion about Solving the Proximal Mapping

Function g	Algorithm
ℓ_1-norm	Separable: exact
Hinge	Separable: exact
ℓ_∞-ball	Separable: exact
Box constraint	Separable: exact
Positivity constraint	Separable: exact
Linear constraint	Nonseparable: exact
ℓ_1-ball	Nonseparable: Semi-smooth Newton + $\mathrm{prox}_{g\circ \boldsymbol{D}^{-1/2}}$ exact
ℓ_∞-norm	Nonseparable: Moreau identity
Simplex	Nonseparable: Semi-smooth Newton + $\mathrm{prox}_{g\circ \boldsymbol{D}^{-1/2}}$ exact

From [Becker, Fadili '12].

Rank-r Modified Metric

- **Rank-r Modified Metric:** (g convex)
  (L-)**BFGS** uses a rank-r update of the metric with $r > 1$.

- **Theorem:** [Becker, Fadili, O. '18]
  $\boldsymbol{Q} = \boldsymbol{P} \pm \boldsymbol{V} \in \mathbb{S}_{++}$, $\boldsymbol{P} \in \mathbb{S}_{++}$, $\boldsymbol{V} = \sum_{i=1}^{r} u_i u_i^\top$, $\mathrm{rank}(\boldsymbol{V}) = r$. Then

$$\mathrm{prox}_g^{\boldsymbol{Q}}(\bar{x}) = \boldsymbol{P}^{-1/2} \circ \mathrm{prox}_{g \circ \boldsymbol{P}^{-1/2}} \circ \boldsymbol{P}^{1/2}(\bar{x} \mp \boldsymbol{P}^{-1}\boldsymbol{U}\alpha^\star)$$

  where $\boldsymbol{U} = (u_1, \dots, u_r)$ and $\alpha^\star$ is the unique root of

$$l(\alpha) = \boldsymbol{U}^\top \Big(\bar{x} - \boldsymbol{P}^{-1/2} \circ \mathrm{prox}_{g \circ \boldsymbol{P}^{-1/2}} \circ \boldsymbol{P}^{1/2}(\bar{x} \mp \boldsymbol{P}^{-1}\boldsymbol{U}\alpha)\Big) + \boldsymbol{X}\alpha\,,$$

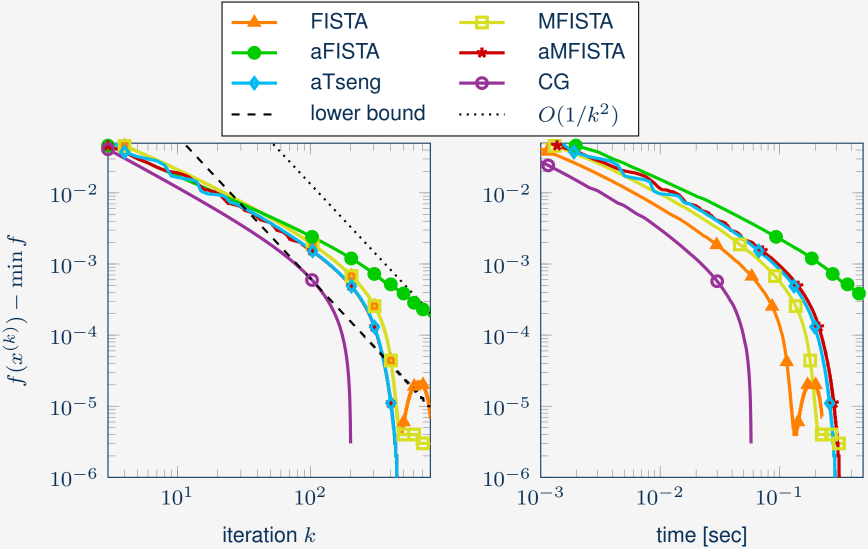
  where $\boldsymbol{X} := \boldsymbol{U}^\top \boldsymbol{V}^+ \boldsymbol{U} \in \mathbb{S}_{++}(r)$. The mapping $l\colon \mathbb{R}^r \to \mathbb{R}^r$ is Lipschitz continuous with constant $\|\boldsymbol{X}\| + \|\boldsymbol{P}^{-1/2}\boldsymbol{U}\|^2$ and strongly monotone.

Variants with $O(1/k^2)$-convergence rate

Adaptive FISTA: Variants with $O(1/k^2)$-convergence rate: (**convex case**)

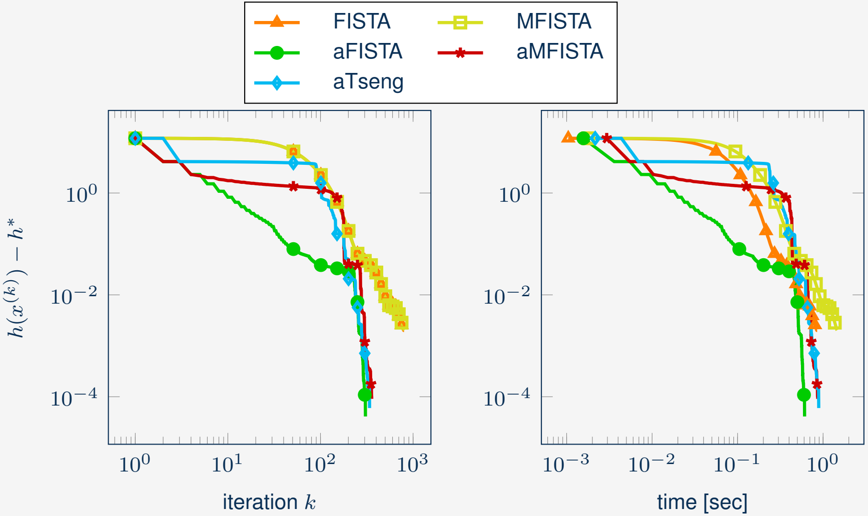
- Adaptive FISTA cannot be proved to have the accelerated rate $O(1/k^2)$.
  - For each point $\bar{x}$, aFISTA decreases the objective more than a FISTA.
  - However, global view on the sequence is lost.
- aFISTA can be embedded into schemes with accelerated rate $O(1/k^2)$.
- **Monotone FISTA version:** (Motivated by [Li, Lin '15], [Beck, Teboulle '09].)
- **Tseng-like version:** (Motivated by [Tseng '08].)

Nesterov's Worst Case Function



LASSO

$$\min_{x\in\mathbb{R}^N} h(x)\,, \quad h(x) = \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1\,,$$



Proposed Algorithm: (**non-convex setting**)

- **Current iterate** $x^{(k)} \in \mathbb{R}^N$. Step size: $\boldsymbol{T} \in \mathbb{S}_{++}(N)$.
- Define the **extrapolated point** $y_\beta^{(k)}$ that depends on β:

$$y_\beta^{(k)} := x^{(k)} + \beta(x^{(k)} - x^{(k-1)}) .$$

- **Exact version:** Compute $x^{(k+1)}$ as follows:

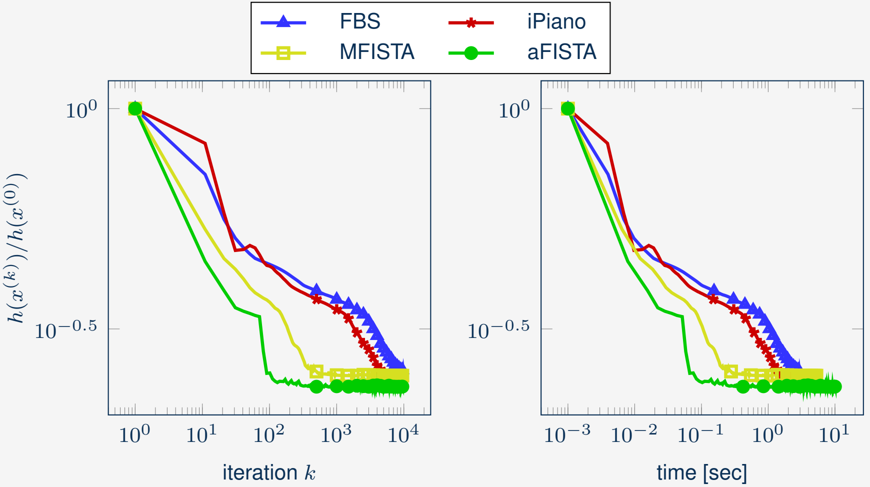
$$x^{(k+1)} = \operatorname*{argmin}_{x\in\mathbb{R}^N} \min_{\beta} \ell_f^g(x; y_\beta^{(k)}) + \frac{1}{2}\|x - y_\beta^{(k)}\|_{\boldsymbol{T}}^2,$$

$$\ell_f^g(x; y_\beta^{(k)}) := g(x) + f(y_\beta^{(k)}) + \left\langle \nabla f(y_\beta^{(k)}), x - y_\beta^{(k)} \right\rangle$$

- **Inexact version**: Find $x^{(k+1)}$ and β such that

$$\ell_f^g(x^{(k+1)}; y_\beta^{(k)}) + \frac{1}{2}\|x^{(k+1)} - y_\beta^{(k)}\|_{\boldsymbol{T}}^2 \leq f(x^{(k)}) + g(x^{(k)})$$

$$\min_{\substack{W_0,W_1,W2\\b_0,b_1,b_2}} \sum_{i=1}^{N} \Big(\|(W_2\sigma_2(W_1\sigma_1(W_0X+B_0)+B_1)+B_2-\tilde{Y})_{1,i}\|^2+\varepsilon^2\Big)^{1/2}+\lambda\sum_{j=0}^{2}\|W_j\|_1$$



Conclusion

Conclusion:

- Proposed **adaptive FISTA** for solving problems of the form

$$\min_{x\in\mathbb{R}^N} \; g(x) + f(x) \,,$$

- Adaptive FISTA is **locally better than FISTA**.
- Prove **convergence to a stationary point**.
- **Equivalence to a proximal quasi-Newton method**, if f is quadratic.
- Often, the proximal mapping can be computed efficiently.
- Adaptive FISTA can be embedded into **accelerated / optimal schemes**.