Iterative Shrinkage/Thresholding Algorithms: Some History and Recent Development

Mário A. T. Figueiredo

Instituto de Telecomunicações
and
Instituto Superior Técnico,
Technical University of Lisbon

PORTUGAL

mario.figueiredo@lx.it.pt

CS Workshop, Duke, 2009

Signal/Image Restoration/Representation/Reconstruction

Many signal/image reconstruction/approximation criteria have the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} \phi(\mathbf{x}) := f(\mathbf{x}) + \tau c(\mathbf{x})$$

$f : \mathbb{R}^n \rightarrow \mathbb{R}$ is smooth and convex (the data fidelity term); usually,

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$$

$c : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ is a regularization/penalty function;

typically convex (sometimes not), often non-differentiable.

Examples: TV-based and wavelet-based restoration/reconstruction,
sparse representations, sparse (linear or logistic) regression,
compressive sensing (with $\mathbf{A} = \mathbf{H}\mathbf{D}$)

Outline

1. The optimization problem (previous slide)
2. IST Algorithms: 4 derivations
3. Convergence results
4. Enhanced (accelerated) versions: TwIST and SpaRSA
5. Warm starting and continuation
6. Concluding remarks

Denoising/shrinkage operators

$$\min_{\mathbf{x}\in\mathbb{R}^n} \phi(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 + \tau c(\mathbf{x})$$

If $\mathbf{A} = \mathbf{I}$, we have a denoising problem.

If c is proper and convex , ϕ is strictly convex, there is a unique minimizer.

Thus, the so-called shrinkage/thresholding/denoising function

$$\Psi_\lambda(\mathbf{u}) = \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z}-\mathbf{u}\|_2^2 + \lambda c(\mathbf{z})$$

is well defined (*Moreau proximal mapping*) [Moreau 1962], [Combettes 2001]

Examples: $c(\mathbf{z}) = \|\mathbf{z}\|_1 \;\Rightarrow\; \Psi_\lambda(\mathbf{z}) = \mathrm{soft}(\mathbf{z},\lambda)$

$c(\mathbf{z}) = \|\mathbf{z}\| \;\Rightarrow\; \Psi_\lambda(\mathbf{z}) = (\mathbf{I} - P_{\lambda S_{c^*}})\mathbf{z}$

(not convex) $c(\mathbf{z}) = \|\mathbf{z}\|_0 \;\Rightarrow\; \Psi_\lambda(\mathbf{z}) = \mathrm{hard}(\mathbf{z},\lambda)$

Iterative Shrinkage/Thresholding (IST)

Problem: $$\min_{\mathbf{x}\in\mathbb{R}^n} \phi(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 + \tau c(\mathbf{x})$$

IST algorithm: $$\mathbf{x}^{k+1} = \Psi_{\tau/\alpha}\left(\mathbf{x}^k - \frac{1}{\alpha}\mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{y})\right)$$

Adequate when products by $\mathbf{A}$ and $\mathbf{A}^T$ are efficiently computable (e.g., FFT)

Since $\mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{y})$ is the gradient of $\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2$

if $\tau = 0$, IST is gradient descent with step length $1/\alpha$

IST also applicale in Bregman iterations to solve constrained problems
[Yin, Osher, Goldfarb, Darbon, 2008]

IST as Expectation-Maximization [F. and Nowak, 2001, 2003]

Underlying observation model: $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Equivalent model: $\mathbf{y} = \mathbf{A}(\mathbf{x} + \mathbf{n}_1) + \mathbf{n}_2, \quad \mathbf{n}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\eta)$

$$\mathbf{n}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{A}\mathbf{A}^T/\eta)$$

Hidden image: $\mathbf{z} = \mathbf{x} + \mathbf{n}_1, \quad p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{z}, \mathbf{I} - \mathbf{A}\mathbf{A}^T/\eta)$

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{x}, \mathbf{I}/\eta)$$

E-step: $\mathbf{z}^k = \mathbb{E}[\mathbf{z}|\mathbf{y}, \mathbf{x}^k] = \mathbf{x}^k + \mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^k)/\eta$ (Wiener)

M-step: $\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \frac{\eta}{2}\|\mathbf{z}^k - \mathbf{x}\|_2^2 + \tau c(\mathbf{x}) = \Psi_{\tau/\eta}(\mathbf{z}^k)$

$\lambda_{\max}(\mathbf{A}^T\mathbf{A}) \leq \eta \;\Rightarrow$ monotonicity

IST as Majorization-Minimization [Daubechies, Defrise, De Mol, 2004]

Majorization function: $\arg\min_{\mathbf{x}} Q(\mathbf{x},\mathbf{y}) - \phi(\mathbf{x}) = \mathbf{y} \qquad (a)$

MM algorithm: $\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} Q(\mathbf{x},\mathbf{x}^k) \qquad (b)$

Monotonicity: $Q(\mathbf{x}^{k+1},\mathbf{x}^k) - \phi(\mathbf{x}^{k+1}) \overset{(a)}{\geq} Q(\mathbf{x}^k,\mathbf{x}^k) - \phi(\mathbf{x}^k)$

$$Q(\mathbf{x}^{k+1},\mathbf{x}^k) \overset{(b)}{\leq} Q(\mathbf{x}^k,\mathbf{x}^k)$$

$$(a) \wedge (b) \Rightarrow \phi(\mathbf{x}^{k+1}) \leq \phi(\mathbf{x}^k)$$

If $\lambda_{\mathbf{max}}(\mathbf{A}^T\mathbf{A}) \leq \gamma$, we can set $Q(\mathbf{x},\mathbf{x}^k) = \frac{\gamma}{2}\|\mathbf{x} - \mathbf{z}^k\|_2^2 + \tau c(\mathbf{x})$

Thus, $\mathbf{x}^{k+1} = \Psi_{\tau/\gamma}(\mathbf{z}^k)$

$$\mathbf{z}^k = \mathbf{x}^k + \mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^k)/\gamma$$

IST as Forward-Backward Splitting

$$\Psi_\tau(\mathbf{u}) = \mathbf{a} \quad \Leftrightarrow \quad \mathbf{a} = \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z}-\mathbf{u}\|_2^2 + \tau c(\mathbf{z})$$

$$\Leftrightarrow \quad \mathbf{0} \in \tau\, \partial c(\mathbf{a}) + (\mathbf{a}-\mathbf{u})$$

$$\Leftrightarrow \quad \mathbf{u} \in (\mathbf{I} + \tau\, \partial c)\mathbf{a}$$

$$\Leftrightarrow \quad \mathbf{a} = (\mathbf{I} + \tau\, \partial c)^{-1}\mathbf{u} = \Psi_\tau(\mathbf{u})$$

(the minimizer is unique)

Back to the problem $\widehat{\mathbf{x}} \in \arg\min_{\mathbf{x}} f(\mathbf{x}) + \tau c(\mathbf{x})$

f differentiable

c convex

$$\Leftrightarrow \quad \mathbf{0} \in \nabla f(\widehat{\mathbf{x}}) + \tau\, \partial c(\widehat{\mathbf{x}}) + (\widehat{\mathbf{x}} - \widehat{\mathbf{x}})\alpha$$

$$\Leftrightarrow \quad (\alpha\mathbf{I} - \nabla f)\widehat{\mathbf{x}} \in (\alpha\mathbf{I} + \tau\, \partial c)\widehat{\mathbf{x}}$$

$$\Leftrightarrow \quad \widehat{\mathbf{x}} \in (\alpha\mathbf{I} + \tau\, \partial c)^{-1}(\alpha\mathbf{I} - \nabla f)\widehat{\mathbf{x}}$$

$$\Leftrightarrow \quad \widehat{\mathbf{x}} = \Psi_{\tau/\alpha}(\widehat{\mathbf{x}} - \nabla f(\widehat{\mathbf{x}})/\alpha)$$

(fixed point equation)

Fixed point scheme: $\mathbf{x}^{k+1} = \Psi_{\tau/\alpha}\left(\widehat{\mathbf{x}}^k - \frac{1}{\alpha}\nabla f(\widehat{\mathbf{x}}^k)\right)$

IST as Separable Approximation

Recall the problem: $$\min_{\mathbf{x}\in\mathbb{R}^n} \phi(\mathbf{x}) := f(\mathbf{x}) + \tau c(\mathbf{x})$$

Separable approximation to $f(\mathbf{z})$

Iteration:

$$\mathbf{x}^{k-1} \in \arg\min_{\mathbf{z}} (\mathbf{z}-\mathbf{x}^k)^T \nabla f(\mathbf{x}^k) - \frac{\alpha_k}{2} \|\mathbf{z}-\mathbf{x}^k\|_2^2 + \tau c(\mathbf{z})$$

Can be re-written as $$\mathbf{x}^{k+1} \in \arg\min_{\mathbf{z}} \frac{\alpha_k}{2} \|\mathbf{z}-\mathbf{z}^k\|_2^2 + \tau\, c(\mathbf{z})$$

If c is convex, $$\mathbf{x}^{k+1} = \Psi_{\tau/\alpha_k}(\mathbf{z}^k) \qquad \mathbf{z}^k = \mathbf{x}^k - \frac{1}{\alpha_k}\nabla f(\mathbf{x}^k)$$

The objective function in each iteration can be seen as the Lagrangian for

$$\mathbf{x}^{k+1} \in \arg\min_{\mathbf{z}} (\mathbf{z}-\mathbf{x}^k)^T \nabla f(\mathbf{x}^k) + \tau c(\mathbf{z})$$
$$\text{subject to } \|\mathbf{z}-\mathbf{x}^k\|_2^2 \le \Delta_t$$

…a trust-region method.

Bibliographical Notes

IST as expectation-maximization: [F. and Nowak, 2001, 2003]

IST as majorization-minimization: [Daubechies, Defrise, De Mol, 2003, 2004]
[F., Nowak, Bioucas-Dias, 2005, 2007]

Forward-backward schemes in math: [Bruck, 1977], [Passty, 1979], [Lions and Mercier, 1979]

Forward-backward schemes in signal reconstruction: [Combettes and Wajs, 2003, 2004]

Separable approximation: [Wright, Nowak, and F., 2008]

Other authors independently proposed IST schemes for signal/image recovery:
[Bect, Blanc-Féraud, Aubert, and Chambolle, 2004],
[Elad, Matalon, and Zibulevsky, 2006],
[Starck, Nguyen, Murtagh, 2003],
[Starck, Candès, Donoho, 2003],
[Hale, Yin, Zhang, 2007]

Existence, Uniqueness

$$G = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 + \tau c(\mathbf{x})$$

G is non empty if c is coercive ($\lim_{\|\mathbf{x}\|\to+\infty} c(\mathbf{x}) = +\infty$)

G has at most one element if c is strictly convex or $\mathbf{A}$ is invertible

G has exactly one element if $\mathbf{A}$ is bounded bellow

[Combettes and Wajs, 2004]

Convergence Results (I)

Problem: $$\min_{\mathbf{x}\in\mathbb{R}^n} \phi(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 + \tau c(\mathbf{x})$$

IST algorithm: $$\mathbf{x}^{k+1} = \Psi_{\tau/\alpha_k}\left(\mathbf{x}^k - \frac{1}{\alpha_k}\mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{y})\right)$$

[Daubechies, Defrise, De Mol, 2004]: (applies in a Hilbert space setting)

Let $c(\mathbf{x}) = \|\mathbf{x}\|_p^p,\;\; p \in [1,\, 2]$, $\alpha_k = 1$, and $\|\mathbf{A}\|_2^2 < 1$; then,

IST converges to a minimizer of ϕ

[Combettes and Wajs, 2005]: (applies to a more general version of IST)

Let c be convex and proper (never $-\infty$, not $+\infty$ everywhere)

and $\frac{\|\mathbf{A}\|_2^2}{2} < \alpha_k < +\infty$; then, IST converges to a minimizer of ϕ

Convergence Results (II)

Problem: $\min_{\mathbf{x}\in\mathbb{R}^n} \phi(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 + \tau c(\mathbf{x})$

IST algorithm: $\mathbf{x}^{k+1} = \Psi_{\tau/\alpha_k}\left(\mathbf{x}^k - \frac{1}{\alpha_k}\mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{y})\right)$

[Hale, Yin, Zhang, 2007]:

Let $c(\mathbf{x}) = \|\mathbf{x}\|_1$ and $\alpha_k > \lambda_{\max}(\mathbf{A}^T\mathbf{A})/2$

Then, IST converges to some $\mathbf{x}^* \in G$ and,

for all but a finite number of iterations:

$$x_i^k = x_i^* = 0, \quad \forall i \in L$$

$$\text{sign}\left((\mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{y}))_i\right) = \text{sign}\left((\mathbf{A}^T(\mathbf{A}\mathbf{x}^* - \mathbf{y}))_i\right), \quad \forall i \in E$$

where $L \cup E = \{1, 2, ..., n\}$

Accelerating IST: Two-Step IST (TwIST)

IST becomes slow when $\mathbf{A}$ is very ill-conditioned and τ is small

Inspired by two-step method for linear systems [Frankel, 1950], [Axelsson, 1996],

TwIST algorithm [Bioucas-Dias and F., 2007]

$$\mathbf{x}^{k+1} = (\alpha - \beta)\mathbf{x}^k + (1-\alpha)\mathbf{x}^{k-1} - \beta\,\mathbf{\Psi}_\tau\left(\mathbf{x}^k + \mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^k)\right)$$

Simplified analysis with $0 < m \leq \lambda_{\min}(\mathbf{A}^T\mathbf{A}) \leq \lambda_{\max}(\mathbf{A}^T\mathbf{A}) = 1$

The minimizer $\widehat{\mathbf{x}}$ is unique and TwIST converges to $\widehat{\mathbf{x}}$, $\lim_{t\rightarrow\infty} \|\mathbf{x}^t - \widehat{\mathbf{x}}\| = 0.$

There is an optimal choice for α and β for which

$$\|\mathbf{x}^{t+1} - \widehat{\mathbf{x}}\| \leq \frac{1-\sqrt{m}}{1+\sqrt{m}}\,\|\mathbf{x}^t - \widehat{\mathbf{x}}\|$$

Accelerating IST: TwIST (II)

A one-step method is recovered for $\alpha = 1$

$$\mathbf{x}^{t+1} = (1-\beta)\mathbf{x}^t + \beta\,\boldsymbol{\Psi}_\lambda\left(\mathbf{x}^t + \mathbf{K}^T(\mathbf{y} - \mathbf{K}\mathbf{x}^t)\right)$$

which is an over-relaxed version of the original IST.

For the optimal choice of β : $\|\mathbf{x}^{t+1} - \widehat{\mathbf{x}}\| \leq \frac{1-m}{1+m}\,\|\mathbf{x}^t - \widehat{\mathbf{x}}\|$

$-1/\log_{10}\frac{1-m}{1+m}$ ~ number of iterations to decrease error by factor of 10.

Example:

$$m = 10^{-3} \quad \rightarrow \quad -1/\log\frac{1-m}{1+m} \sim 1150 \qquad\qquad -1/\log\frac{1-\sqrt{m}}{1-\sqrt{m}} \sim 35$$

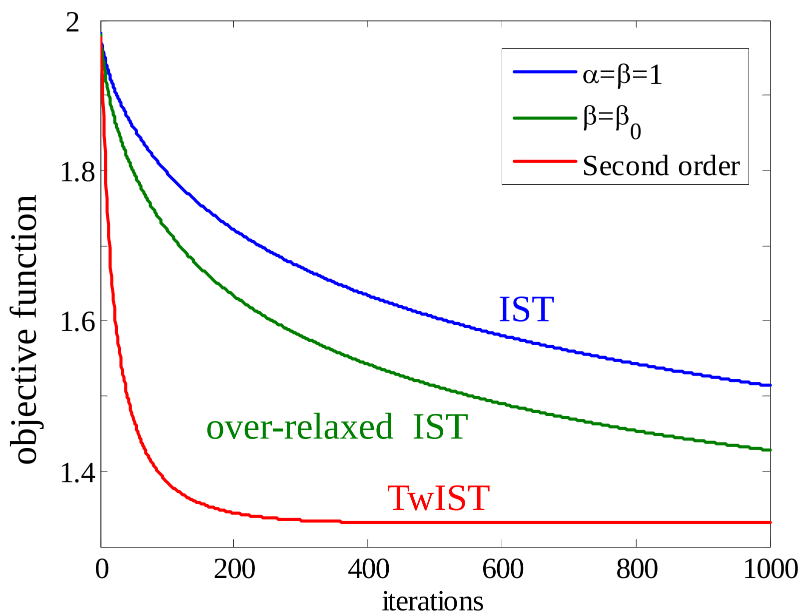
Another two-step method was recently proposed in [Beck and Teboulle, 2008]

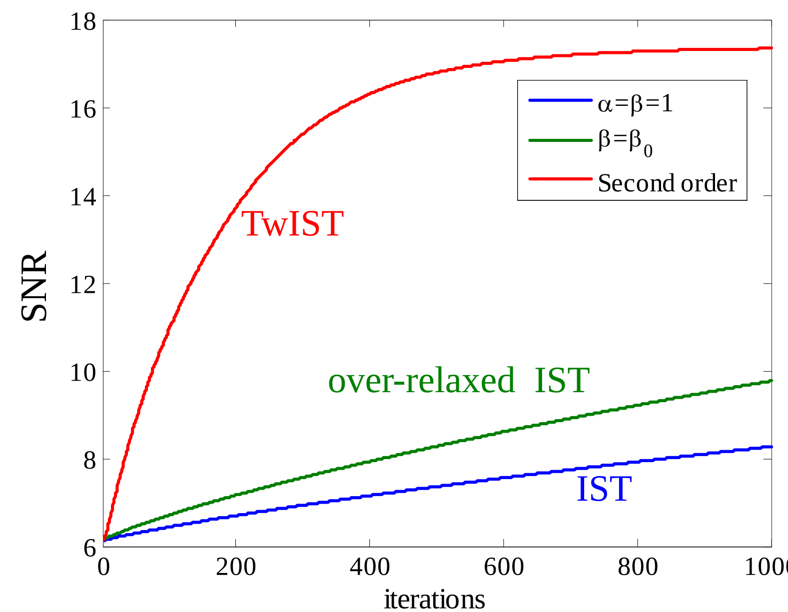
Accelerating IST: TwIST (III)

original

Blurred, 9x9, 40db noise

restored





Accelerating IST: The SpaRSA Algorithmic Framework

Initialization: choose $\eta > 1$, $\alpha_{\min} \ll \alpha_{\max}$, and $\mathbf{x}^0$; set $k \leftarrow 0$

repeat:

choose $\alpha_k \in [\alpha_{\min}, \alpha_{\min}]$

repeat:

$$\mathbf{x}^{k+1} \leftarrow \Psi_{\tau/\alpha_k}\left(\mathbf{x}^k - \frac{1}{\alpha_k}\nabla f(\mathbf{x}^k)\right)$$

$$\alpha_k \leftarrow \eta\, \alpha_k$$

until $Acc(\mathbf{x}^{k+1}) == 1$ (* acceptance criterion *)

$k \leftarrow k + 1$

until stopping criterion is satisfied.

[Wright, Nowak, F., 2008]

Variants of SpaRSA are distinguished by the choice of α_k, Ψ_λ, and Acc

Examples: $Acc = 1,\ \alpha_k = \alpha$ yields standard IST.

$Acc(\mathbf{x}^{k+1}, \mathbf{x}^k) = 1_{\phi(\mathbf{x}^{k+1}) < \phi(\mathbf{x}^k)}$ yields monotone SpaRSA

Choosing α_k for Speed

The Barzilai-Borwein approach: seek α_k to mimic a Newton step,

a less conservative choice than in IST:

$$\alpha_k \, \mathbf{I} \simeq \nabla^2 f(\mathbf{x})$$

With a least-squares criterion over the last step,

$$\alpha_k = \arg\min_{\alpha} \left\| \alpha(\mathbf{x}^k - \mathbf{x}^{k-1}) - (\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}) \right\|_2^2$$

If $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$, then $\alpha_k = \dfrac{\|\mathbf{A}(\mathbf{x}^k - \mathbf{x}^{k-1})\|_2^2}{\|\mathbf{x}^k - \mathbf{x}^{k-1}\|_2^2}$

Alternative rule (SpaRSA-monotone): $\alpha_k = \beta\, \alpha_{k-1}$, with $\beta < 1$

Compressed Sensing Experiment

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \qquad c(\mathbf{x}) = \|\mathbf{x}\|_1$$

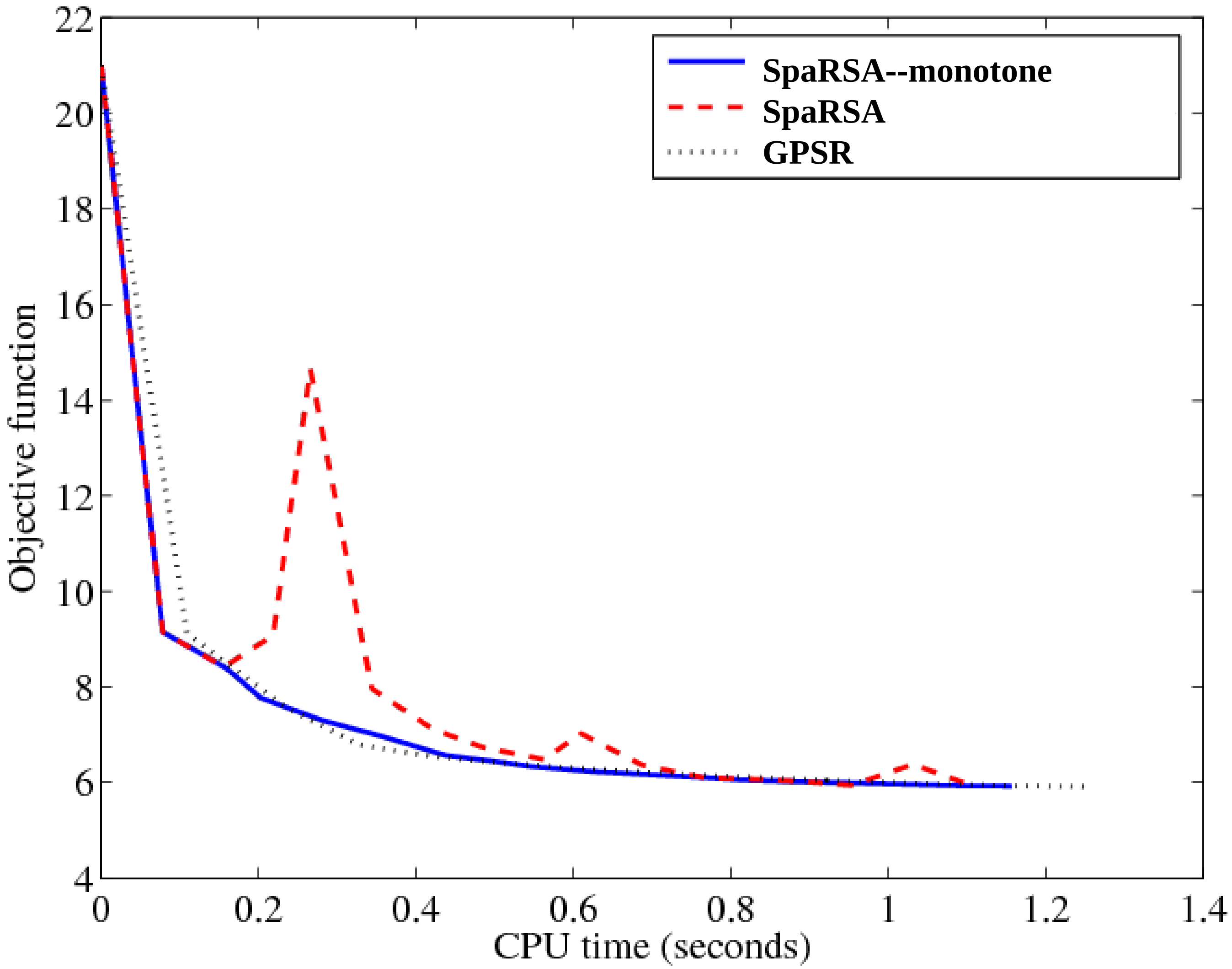
$\mathbf{A}$ 2^{10} x 2^{12} random (Gaussian), $\mathbf{x}$ 160 randomly located non-zeros

$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where $\mathbf{e} \sim \mathcal{N}(0, 10^{-4})$

Reference	Algorithm	CPU time (secs.)	MSE
	SpaRSA	0.33	2.89e-3
	SpaRSA-monotone	0.34	2.91e-3
[F., Nowak, Wright, 2007]	GPSR-BB-monotone	0.42	2.92e-3
	GPSR-Basic	0.67	2.93e-3
[Hale, Yin, Zhang, 2007]	FPC	1.55	2.95e-3
[Kim, Koh, Lustig, Boyd, Gorinvesky, 2007]	l1_ls	9.80	2.96e-3
[Nesterov, 2007]	AC	2.83	2.91e-3
[Bioucas-Dias, F., 2007]	TwIST	0.63	2.91e-3

GPSR and *l1_ls* are “hardwired” for $c(\mathbf{x}) = \|\mathbf{x}\|_1$

Non-monotonicity



Convergence of SpaRSA

Problem: $$\min_{\mathbf{x}\in\mathbb{R}^n} \phi(\mathbf{x}) := f(\mathbf{x}) + \tau c(\mathbf{x})$$

Critical point $\bar{\mathbf{x}}$ if $\mathbf{0} \in \partial\phi(\bar{\mathbf{x}}) = \nabla f(\bar{\mathbf{x}}) + \tau \partial c(\bar{\mathbf{x}})$

Criticality is necessary for optimality.
If both c and f are convex, it is also sufficient.

Safeguarded SpaRSA (S-SParRSA) [Wright, Nowak, F., 2008]

$$Acc(\mathbf{x}^{k+1}) = 1 \Leftrightarrow \phi(\mathbf{x}^{k+1}) \leq \max_{t=k-M,\ldots,k} \phi(\mathbf{x}^t) - \frac{\sigma\,\alpha_t}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2$$

where $\sigma \in]0,\, 1[$, usually $\sigma \ll 1$, e.g., $\sigma = 10^{-5}$

Let f be Lipschitz continuously differentiable, c convex and finite-valued, and ϕ bounded below. Then, all accumulation points of S-SpaRSA are critical points of ϕ

Warm Starting and Continuation

SpaRSA (as GPSR, IST, etc) is slow for small τ

SpaRSA (as GPSR and IST) is “warm-startable”,

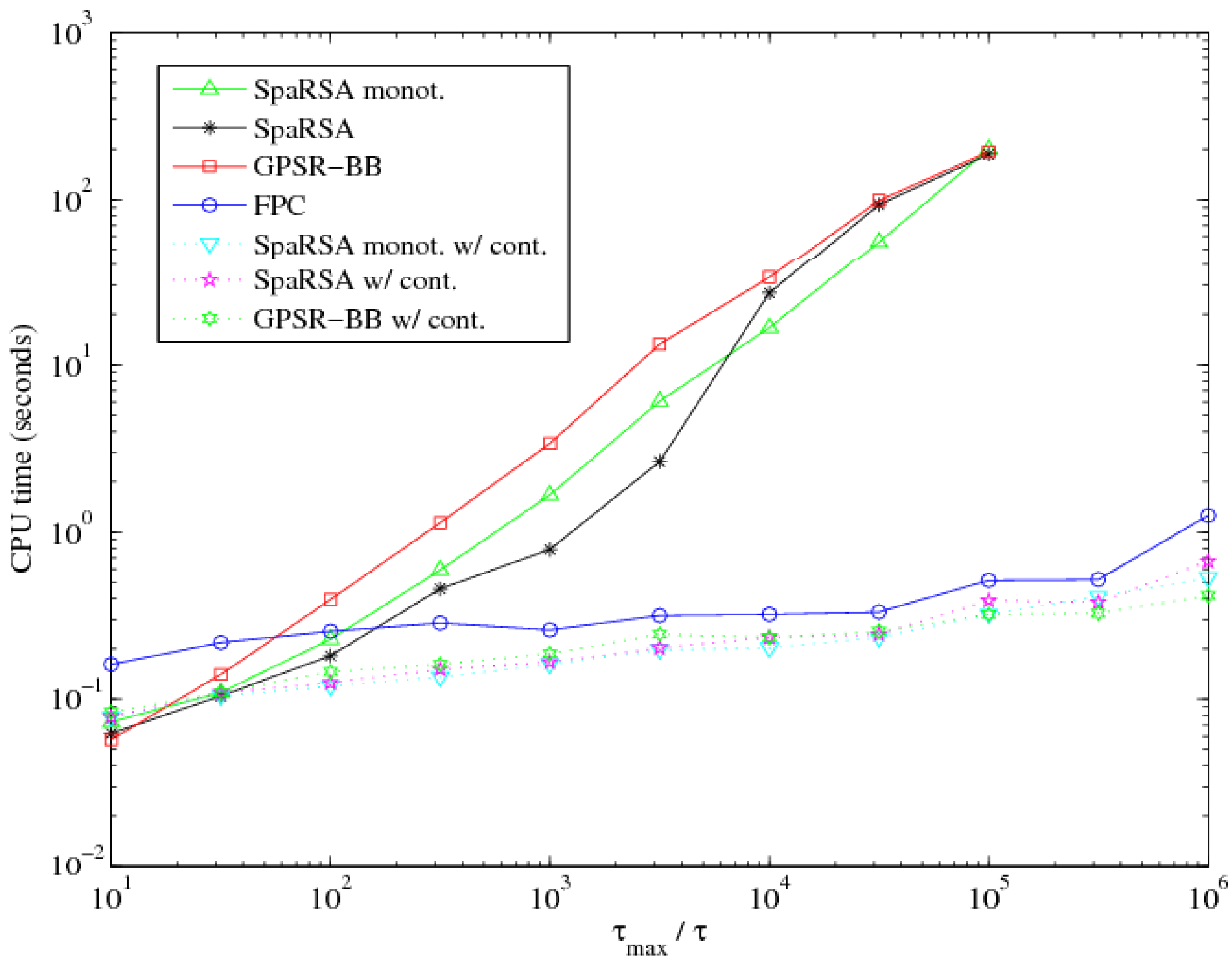
i.e., it benefits (a lot) from a good initialization.

Continuation scheme: start with large τ

slowly decrease τ while tracking the solution.

IST + continuation = fixed point continuation (FPC) [Hale, Yin, Zhang, 2007]

Continuation Experiment



$$\tau_{\mathbf{max}} = \|\mathbf{A}^T\mathbf{y}\|_\infty$$

For $\tau \geq \tau_{\mathbf{max}}$, the solution is the zero vector

Conclusions

- Reviewed several ways to derive the IST algorithm

- Reviewed several convergence results for IST

- Described recent accelerated versions: TwIST, SpaRSA

- IST and SpaRSA benefits (a lot) from a continuation scheme.

-State-of-the-art performance for a variety of problems:
MRI reconstruction (TV and wavelets), MEG imaging, deconvolution, compressed sensing, …