NESTA:
a fast and accurate first-order method for sparse recovery

S. Becker, J. Bobin, E. Candès
Applied and Computational Mathematics 2009

Presented by
Federico Pierucci

LJK
Grenoble, April 24, 2014

Introduction

- Algorithm NESTA
- Optimization based on Nesterov's method
- Compressed sensing applications (e.g. sparse recovery, Total Variation minimization)
- Accurate retrieval of the signal
- Large scale - e.g. $x \in \mathbb{R}^n$ with, $n = 262\,144$

$$
(\mathrm{BP}_\epsilon) \qquad \begin{array}{ll} \text{minimize} & \|x\|_{\ell_1} \\ \text{subject to} & \|b - Ax\|_{\ell_2} \leq \epsilon, \end{array}
$$

A consequence of these properties is that NESTA may be interest of researchers working on signal recovery or undersampled data

Optimization

Problem to solve with NESTA:

$$(\mathrm{BP}_\epsilon) \quad \text{minimize} \quad \|x\|_{\ell_1} \quad \text{subject to} \quad \|b - Ax\|_{\ell_2} \le \epsilon,$$

Equivalent formulation:

$$(\mathrm{QP}_\lambda) \quad \text{minimize} \quad \lambda\|x\|_{\ell_1} + \frac{1}{2}\|b - Ax\|_{\ell_2}^2,$$

$\|\cdot\|_{\ell_1} \rightarrow$ a sparse solution; $\epsilon^2 =$ estimated bound on noise

where
$b = Ax^0 + z$: collected data
x^0: signal to recover
A : sampling matrix
z : noise

source: G. Peyré OSL2013

Nesterov's method

$$\min_{x \in \mathcal{Q}_p} f(x)$$

f : smooth, i.e. differentiable with Lipschitz gradient
$\mathcal{Q}_p$: convex set

Lipschitz gradient of f, with lipschitz constatant L:

$$\forall x, y \in \mathcal{Q}_p \quad \|\nabla f(y) - \nabla f(x)\|_{\ell_2} \leq L \|y - x\|_{\ell_2}$$

[Y. NESTEROV, *Smooth minimization of nonsmooth functions*, Math. Program. (2005)]

Nesterov's method

Initialize x_0. For $k \geq 0$,

1. Compute $\nabla f(x_k)$.

2. Compute y_k:

$$y_k = \text{argmin}_{x \in Q_p} \frac{L}{2}\|x - x_k\|_{\ell_2}^2 + \langle \nabla f(x_k), x - x_k \rangle.$$

3. Compute z_k:

$$z_k = \text{argmin}_{x \in Q_p} \frac{L}{\sigma_p} p_p(x) + \sum_{i=0}^{k} \alpha_i \langle \nabla f(x_i), x - x_i \rangle.$$

4. Update x_k:

$$x_k = \tau_k z_k + (1 - \tau_k) y_k.$$

Stop when a given criterion is valid.

p_p: continuous and strongly convex , $p_p(x) \geq \frac{\sigma_p}{2} \left\| x - x_p^c \right\|^2$

What if
f
is nonsmooth?

Smoothing

We can rewrite the norm as support function

$$\|x\|_{\ell_1} = \max_{u \in \mathcal{Q}_d} \langle u, x \rangle,$$

where

$$\mathcal{Q}_d = \{u : \|u\|_\infty \leq 1\}.$$

The smoothed version of $\|\cdot\|_{\ell_1}$ is

$$f_\mu(x) = \max_{u \in \mathcal{Q}_d} \langle u, x \rangle - \mu \, p_d(u),$$

Theorem

If p_d is continuous and strongly convex on $\mathcal{Q}_d$, then f_μ is smooth.

Then it is possible to apply Nesterov method to f_μ

[Y. NESTEROV, Smooth minimization of nonsmooth functions, Math. Program. (2005)]

NESTA = Nesterov method + smoothing

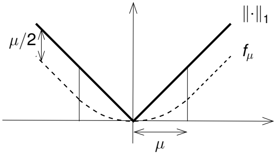
Convergence of NESTA:

$$f_\mu(y_k) - f_\mu(x^\star_\mu) \le \frac{2L_\mu \|x^\star_\mu - x_0\|^2_{\ell_2}}{k^2},$$

k : iteration counter
$x^\star_\mu := \mathrm{argmin}_{x \in Q_p} f_\mu(x)$
L_μ : Lipschitz constant of f_μ

The parameter μ controls the smoothing

$$f_\mu(x) = \max_{u \in \mathcal{Q}_d} \langle u, x \rangle - \mu \, p_d(u),$$



Source: Pierucci, Harchaoui, Malick, tech. report 2014

Small $\mu \to$ good approximation, slow convergence
Large $\mu \to$ worst approximation, faster convergence

Why don't we start with a large μ

and continue with a smaller μ
?

NESTA “with continuation”

Initialize μ_0, x_0 and the number of continuation steps T. For $t \geq 1$,

1. Apply Nesterov's algorithm with $\mu = \mu^{(t)}$ and $x_0 = x_{\mu^{(t-1)}}$.
2. Decrease the value of μ: $\mu^{(t+1)} = \gamma\mu^{(t)}$ with $\gamma < 1$.

Stop when the desired value of μ_f is reached.

Convergence of NESTA with continuation

THEOREM 3.1. *At each continuation step t, $\lim_{k\to\infty} y_k = x^\star_{\mu^{(t)}}$, and*

$$f_{\mu^{(t)}}(y_k) - f_{\mu^{(t)}}(x^\star_{\mu^{(t)}}) \leq \frac{2L_{\mu^{(t)}} \|x^\star_{\mu^{(t)}} - x_{\mu^{(t-1)}}\|^2_{\ell_2}}{k^2}.$$

Accuracy evaluation

Analytical solution only for particular cases $\to$ FISTA

$$(\mathrm{QP}_\lambda) \quad \text{minimize} \quad \lambda\|x\|_{\ell_1} + \frac{1}{2}\|b - Ax\|_{\ell_2}^2,$$

Relative error on objective function

$$\frac{\|x\|_{\ell_1} - \|x^\star\|_{\ell_1}}{\|x^\star\|_{\ell_1}}$$

Accuracy of optimal solution

ℓ_∞ error $:= \|x - x^\star\|_{\ell_\infty}$

TABLE 4.2

NESTA's accuracy. The errors and number of function calls $\mathcal{N}_A$ have the same meaning as in Table 4.1.

Method	ℓ_1-norm	Rel. error ℓ_1-norm	ℓ_∞ error	$\mathcal{N}_A$
FISTA	5.71539e+7			
NESTA $\mu = 0.2$	5.71614e+7	1.3e-4	3.8	659
NESTA $\mu = 0.02$	5.71547e+7	1.4e-5	0.96	1055
NESTA $\mu = 0.002$	5.71540e+7	1.6e-6	0.64	1537

$x^\star$: optimal solution for BP_ϵ

Results

TABLE 5.2

Number of function calls $\mathcal{N}_A$ averaged over 10 independent runs. The sparsity level $s = m/5$ and the stopping rule is Crit. 2 (5.2).

Method	20 dB	40 dB	60 dB	80 dB	100 dB
NESTA	446 351/491	880 719/951	1701 1581/1777	4528 4031/4749	14647 7729/15991
NESTA + Ct	479 475/485	551 539/559	605 589/619	658 635/679	685 657/705
GPSR	59 44/64	736 678/790	5316 4814/5630	DNC	DNC
GPSR + Ct	305 293/311	251 245/257	511 467/543	1837 1323/2091	9127 7251/10789
SpaRSA	345 327/373	455 435/469	541 509/579	600 561/629	706 667/819
SPGL1	55 37/61	138 113/152	217 196/233	358 300/576	470 383/568
FISTA	65 63/66	288 279/297	932 882/966	3407 2961/3591	13160 11955/13908
FPC AS	176 169/183	236 157/263	218 215/239	344 247/459	330 319/339
FPC AS (CG)	357 343/371	475 301/538	434 423/481	622 435/814	588 573/599
FPC	416 398/438	435 418/446	577 558/600	899 788/962	3866 1938/4648
FPC-BB	149 140/154	172 164/174	217 208/254	262 248/286	512 308/790
Bregman-BB	211 203/225	270 257/295	364 355/393	470 429/501	572 521/657

Dynamic range of a signal x is $\log_{10}\left(\frac{x_{max}}{x_{min}}\right)$, measured in becibel

Conclusion

$$
(\mathrm{BP}_\epsilon) \qquad \begin{array}{ll} \text{minimize} & \|x\|_{\ell_1} \\ \text{subject to} & \|b - Ax\|_{\ell_2} \leq \epsilon, \end{array}
$$

- Nesterov's method
- NESTA
- NESTA with continuation
- Comparison with FISTA
- Compressed sensing applications
- Accurate retrieval of the signal
- Large scale

Thank you for your attention

Observation on NESTA with continuation:

Convergence

THEOREM 3.1. *At each continuation step* t, $\lim_{k\to\infty} y_k = x^\star_{\mu^{(t)}}$, *and*

$$f_{\mu^{(t)}}(y_k) - f_{\mu^{(t)}}(x^\star_{\mu^{(t)}}) \leq \frac{2L_{\mu^{(t)}} \|x^\star_{\mu^{(t)}} - x_{\mu^{(t-1)}}\|^2_{\ell_2}}{k^2}.$$

$\gamma < 1$
L_μ is proportional to $\frac{1}{\mu}$.

If we take $\mu_{(t)} = \gamma^t \mu_0$ we have $\frac{L_{\mu(t)}}{k^2} \propto \frac{1}{\gamma^t k^2}$. Then the Lipschitz constant grows faster than k^2. If $t(k) = k$ there is no evident convergence. We conclude that the convergence proof is valid only if the decreasing value of μ is lower bounded.