Iterated smoothing for accelerated gradient convex minimization in signal processing

Jensen, Tobias Lindstrøm; Østergaard, Jan; Jensen, Søren Holdt

Published in:
I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2010.5494983

Publication date:
2010

Document Version
Early version, also known as pre-print

ITERATED SMOOTHING FOR ACCELERATED GRADIENT CONVEX MINIMIZATION IN SIGNAL PROCESSING

Tobias Lindstrøm Jensen, Jan Østergaard, Søren Holdt Jensen

Aalborg University
Department of Electronic Systems
Niels Jernesvej 12, 9220 Aalborg Ø, Denmark

ABSTRACT

In this paper, we consider the problem of minimizing a non-smooth convex problem using first-order methods. The number of iterations required to guarantee a certain accuracy for such problems is often excessive and several methods, *e.g.*, restart methods, have been proposed to speed-up the convergence. In the restart method a smoothness parameter is adjusted such that smoother approximations of the original non-smooth problem are solved in a sequence before the original, and the previous estimate is used as the starting point each time. Instead of adjusting the smoothness parameter after each restart, we propose a method where we modify the smoothness parameter in each iteration. We prove convergence and provide simulation examples for two typical signal processing applications, namely total variation denoising and ℓ_1-norm minimization. The simulations demonstrate that the proposed method require fewer iterations and show lower complexity compared to the restart method.

***Index Terms*—** convex optimization, first-order optimization methods, smoothing techniques, restart, continuation.

1. INTRODUCTION

Recently there has been a renewed interest in optimal first-order methods even though these methods have been known for some time [1, 2], see also [3] for a unified framework. The inspiration for the current interest in first-order methods appears to come from a recent method that guarantee linear complexity for non-smooth problems with certain structures [4].

The motivation for using first-order methods is usually in the case of large scale problems, where second-order methods might scale poorly or problems where moderate accuracy of the solution is sufficient. Such problems occurs in image processing [5, 6], but also compressed sensing recovery applies first-order methods [7–9]. These methods have also been used for robust numerical software packages [9, 10].

One method to minimize a non-smooth function is by minimizing a smooth approximation of the original non-smooth function. The effectiveness of such an approach is dependent upon the choice of a smoothness parameter, which also determines the accuracy of the smooth approximation. A large smoothness parameter yields a very smooth problem and in the early iterations of the algorithm, the function value will quickly decrease. However, the algorithm might not converge because the smooth approximation is not accurate enough. On the other hand, a sufficiently small smoothness parameter, gives a less smooth but more accurate approximation. In this case the function value will slowly decrease but convergence within the required accuracy is guaranteed. To decrease the number of iterations, and thereby speed up the algorithm, one may use a restart methods. The idea is to combine the fast decreasing function value in the early iterations for a very smooth problem with a sufficiently well approximated smooth function to ensure convergence in the final iterations. The algorithm starts by solving a much smoother problem than the original problem and then subsequently solve lesser smooth problems, using the previous estimate as the starting point at each restart, see [9, 11] and references therein. Such an approach is considered a heuristic except for the case of strongly convex functions where there are interesting theoretical results [11], or [12, §5.1.2] for composite objective functions.

In this paper, we will consider convex (but not strongly convex) non-smooth functions. For this case the results indicate that continuation or restart are practical efficient methods to decrease the number of iterations and yet reach the required accuracy [9]. We first review the restart method [9, 11] and relate this approach to the continuation method, see [7, 13] and references therein. We also demonstrate via simulations that restart methods reduce the complexity compared to an approach with a fixed smoothness parameter. Then, inspired by [7, 9, 11, 13] we propose a new method where we decrease the smoothness parameter in each iteration and prove that it converges. Our bound is, however, loose and the actual complexity is in practice much better than what the bound suggests. Simulation examples for two typical signal processing applications, namely total variation denoising and ℓ_1-norm minimization, show that the proposed method yield lower complexity compared to both the fixed smoothing approach and the restart approach.

2. A SMOOTHING METHOD

Let us consider the following optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q_{\mathrm{p}} \end{array} \tag{1}$$

with the dual problem

$$\begin{array}{ll} \text{maximize} & g(u) \\ \text{subject to} & u \in Q_{\mathrm{d}} \end{array} \tag{2}$$

where f is a non-smooth, non-strongly, convex function and Q_{p}, Q_{d} are convex sets. Let x^* and u^* be solutions to the problems (1)

The work is part of the project CSI: Computational Science in Imaging, supported by grant no. 274-07-0065 from the Danish Agency for Science Technology and Innovation. The work of Jan Østergaard is supported by the Danish research council for Technology and Production Sciences, grant no. 274-07-0383.

Emails: T.L. Jensen (tlj@es.aau.dk), J. Østergaard (jo@es.aau.dk) and S. H. Jensen (shj@es.aau.dk).

and (2), respectively. The complexity estimate for problem (1) is $\mathcal{O}(1/\epsilon^2)$, where ϵ is the accuracy of the objective value

$$f(x) - f(x^*) \leq f(x) - g(u) \leq \epsilon, \quad x \in Q_{\rm p},\, u \in Q_{\rm d}\,.$$

In [4] it was, however, shown that for problems with certain structures it is possible to obtain the complexity $\mathcal{O}(1/\epsilon)$, which is one order faster than the sub-gradient method. The idea is to exploit the structure of the non-smooth problem. This is done by making a smooth approximation of the non-smooth function and then subsequently minimize the smooth approximation using an optimal first-order method for the class of smooth problems.

In the following we review the steps required for approximating a non-smooth function by a smooth function. A more general approach is given in [4], but this reduced form will be sufficient for our simulations in Sec. 5. Let the function f have the form

$$f(x) = \max_{u \in Q_{\rm d}} u^T A x\,, \tag{3}$$

where we now assume $Q_{\rm d}$ is a closed and bounded convex set. We then approximate f by f_μ where

$$f_\mu(x) = \max_{u \in Q_{\rm d}} u^T A x - \mu d_{\rm d}(u)\,,$$

with $\mu > 0$ called the smoothness parameter and $d_{\rm d}(u) \geq \frac{1}{2}\|u - \hat{u}\|_2^2$. The function f_μ satisfy

$$f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu\Delta_{\rm d}\,, \quad \Delta_{\rm d} = \max_{u \in Q_{\rm d}} d_{\rm d}(u)\,. \tag{4}$$

The approximation function is also smooth, *i.e.*, it has Lipschitz continuous gradient

$$\|\nabla f_\mu(x) - \nabla f_\mu(\tilde{x})\|_2 \leq L_\mu \|x - \tilde{x}\|_2, \quad x, \tilde{x} \in Q_{\rm p}$$

with

$$L_\mu = \frac{\|A\|_2^2}{\mu}\,. \tag{5}$$

It was shown in [4] that the optimal selection of a fixed μ for achieving an ϵ-accuracy is

$$\mu = \frac{\epsilon}{2\Delta_{\rm d}}\,, \tag{6}$$

which results in an $\epsilon/2$ approximation of f, *i.e.*,

$$f_\mu(x) \leq f(x) \leq f_\mu(x) + \frac{\epsilon}{2}\,. \tag{7}$$

We now apply the smooth approximation to an optimal first-order method with complexity estimate $\mathcal{O}(\sqrt{L/\epsilon})$ [14] where L is the Lipschitz constant of the gradient of the objective function. Using (5) and (6) we obtain the complexity $\mathcal{O}(1/\epsilon)$ for non-smooth problems.

3. RESTART

As indicated in (6), a fixed μ is selected so small that the approximation accuracy in (7) would be smaller than the required accuracy. Another approach is to select μ large in the early iterations because the smooth approximation converges like $\mathcal{O}(\sqrt{L_\mu/\epsilon})$ and then in the final iterations select μ small enough to ensure the smooth approximation comes within the required accuracy. This idea is used in [9, 11], where the main algorithm is restarted several times with first a large μ, and then subsequently a smaller and smaller μ. Note that in [7, 13], they solve composite problems of the form

$$\min_x \psi(x) + \frac{1}{\mu} h(x)\,,$$

where $h(x)$ is smooth and $\psi(x)$ is non-smooth. The smoothness, or the Lipschitz constant of $\frac{1}{\mu}\nabla h(x)$, is $L = \frac{1}{\mu}L(\nabla h(x))$ where $L(f)$ is the Lipschitz constant of the function f. For small μ we will then have a large Lipschitz constant of the gradient function. The continuation idea in [7,13] is then similar to the restart approaches in [9, 11] because the sequence of problems solved in the continuation strategy becomes less and less smooth, as in the restart approach.

For strongly convex functions, it is possible to guarantee that the previous estimate is useful as a starting point, *i.e.*, warm start, and then show the advantage of applying a restart method. Let ϕ be a strongly convex max-type function with strong convexity parameter σ, but ϕ does not have a Lipschitz continuous gradient. We then have [14, Corollary 2.3.1]

$$\frac{\sigma}{2}\|y - y^*\|_2^2 \leq \phi(y) - \phi(y^*), \quad y \in Q$$

where y^* is the solution that minimize $\phi(y)$ for $y \in Q$. It was shown in [11] that the restart algorithm has the complexity $\mathcal{O}(1/\log(\epsilon))$ for strongly convex non-smooth functions. Warm start approaches for first-order methods are also studied in [15] and in [12, §5.1.2]. The restart algorithm from [11] is given below.

Algorithm: **Restart** [11]

Given a $\bar{x}^{(0)}$,$\bar{u}^{(0)}$, $\gamma > 0$, $k = 0$ and ϵ

Repeat for $j = 0, 1, 2, \ldots$

$\bar{\epsilon}_j = \max\left(\frac{f(\bar{x}^{(j)}) - g(\bar{u}^{(j)})}{\gamma}, \epsilon\right)$

$\bar{x}^{(j+1)}, \bar{u}^{(j+1)}, \bar{k}^{(j+1)} =$ **NESTEROV**$(\bar{x}^{(j)}, \bar{\epsilon}_j)$

$k = k + \bar{k}^{(j+1)}$

if $f(\bar{x}^{(j+1)}) - g(\bar{u}^{(j+1)}) \leq \epsilon$ **then break**

The function **NESTEROV** is not shown, but is the algorithm presented in [4, §3.11], which outputs a primal and dual $\bar{\epsilon}_j$-optimal solution after $\bar{k}^{(j+1)}$ iterations with the starting point $\bar{x}^{(j)}$ and using the smoothness parameter $\mu = \frac{\bar{\epsilon}_j}{2\Delta_{\rm d}}$.

4. ITERATED SMOOTHING

In the previous section we reviewed a restart algorithm where the smoothness parameter was decreased before a restart. The idea proposed in this section is to decrease the smoothness parameter in each iteration instead of only after each restart, using an optimal first-order method as base.

We will study the convergence properties of such an algorithm. Let $\{(x^{(j)}, y^{(j)}, z^{(j)}, \theta_j)\}$ be generated by Algorithm 1 or Algorithm 2 from [3], and use the smooth approximation $f_{\mu_j}(x)$ (with a variable smoothness parameter μ_j). We then have

$$\begin{aligned} f_{\mu_j}(x^{(j+1)}) \leq\; & (1 - \theta_j) f_{\mu_j}(x^{(j)}) + \theta_j f_{\mu_j}(x^*) \\ & + \theta_j^2 L_{\mu_j}(\tfrac{1}{2}\|x^* - z^{(j)}\|_2^2 - \tfrac{1}{2}\|x^* - z^{(j+1)}\|_2^2) \end{aligned}$$

for the iterations $j = 0, 1, \ldots$ Using the approximation in (4), we obtain

$$\begin{aligned} f(x^{(j+1)}) - \mu_j\Delta_{\rm d} \leq\; & (1 - \theta_j) f(x^{(j)}) + \theta_j f(x^*) \\ & + \theta_j^2 L_{\mu_j}(\tfrac{1}{2}\|x^* - z^{(j)}\|_2^2 - \tfrac{1}{2}\|x^* - z^{(j+1)}\|_2^2). \end{aligned}$$

With $\theta_k = \frac{2}{k+2}$, we select $\mu_j = \alpha\theta_j^2$ as a quadratically decreasing function. This will ensure that the approximation error converges to a constant. We then obtain

$$f(x^{(j+1)}) - f(x^*) - (1-\theta_j)(f(x^{(j)}) - f(x^*)) \leq \frac{\|A\|_2^2}{\alpha}\left(\tfrac{1}{2}\|x^* - z^{(j)}\|_2^2 - \tfrac{1}{2}\|x^* - z^{(j+1)}\|_2^2\right) + \theta_j^2\alpha\Delta_{\mathrm{d}}\,.$$

Adding the inequalities from $j = 0, 1, \ldots, k-1$, gives

$$f(x^{(k)}) - f(x^*) + \sum_{j=1}^{k-1}\theta_j(f(x^{(j)}) - f(x^*)) \leq \frac{\|A\|_2^2}{\alpha}(\tfrac{1}{2}\|x^* - z^{(0)}\|_2^2 - \tfrac{1}{2}\|x^* - z^{(k)}\|_2^2) + \sum_{j=0}^{k-1}\theta_j^2\alpha\Delta_{\mathrm{d}}\,.$$

We then obtain the lower bound

$$f(x^{(k)}) - f(x^*) + \sum_{j=1}^{k-1}\theta_j(f(x^{(j)}) - f(x^*)) \geq \min_{i=1,\ldots,k}\left\{f(x^{(i)}) - f(x^*)\right\}\left(1 + \sum_{j=1}^{k-1}\theta_j\right).$$

For $\theta_k = \frac{2}{k+2}$, we have

$$\sum_{j=1}^{k-1}\theta_j \geq 2\log_{\mathrm{e}}(k+1) - 3, \quad \sum_{j=0}^{k-1}\theta_j^2 \leq \sum_{j=0}^{\infty}\theta_j^2 = \frac{2}{3}\pi^2 - 4\,.$$

It is important that the sum of the approximation errors is bounded by a constant. This is achieved for quadratically decreasing functions, which motivated our selection $\mu_j = \alpha\theta_j^2$. For $k \geq 2$,

$$\min_{i=1,\ldots,k}\left\{f(x^{(i)}) - f(x^*)\right\} \leq \frac{1}{2\log_{\mathrm{e}}(k+1) - 2}\left(\frac{\|A\|_2^2}{2\alpha}\|x^* - x^{(0)}\|_2^2 + \alpha\Delta_{\mathrm{d}}\left(\frac{2}{3}\pi^2 - 4\right)\right).$$

The algorithm converges, although the upper bound decreases slowly. The parameter α works as a tradeoff between the two terms in the brackets. Since $\|x^* - x^{(0)}\|_2^2$ is unknown in practice and the bound above is loose, we are instead inspired by (6) and set

$$\alpha = \frac{f(x^{(0)}) - g(u^{(0)})}{2\Delta_{\mathrm{d}}c}\,,$$

where c is a scaling reflecting that $g(u^{(0)})$ might severely underestimate $f(x^*)$. The algorithm **Smooth** implements the iteratively decreasing smoothness parameter studied in this section and is applied to Algorithm 1 in [3], with the smoothing technique presented in [4]. The function $P_Q(x)$ is the projection of x onto Q,

$$P_Q(x) = \operatorname*{argmin}_{y\in Q}\|x - y\|_2^2.$$

5. SIMULATIONS

In this section, we compare the three algorithms, **Fixed** (as in [4, §3.11] with a fixed μ selection), **Restart** and **Smooth** for solving two different problems on the form (1) and (3). For algorithms **Fixed** and **Smooth** we record the number of iterations k required to reach the duality gap

$$f(x^{(k)}) - g(u^{(k)}) \leq \epsilon, \quad x^{(k)} \in Q_{\mathrm{p}},\, u^{(k)} \in Q_{\mathrm{d}}$$

Algorithm: **Smooth**

Given a $x^{(0)}$, $u^{(0)}$ and ϵ, set $\alpha = \frac{f(x^{(0)}) - g(u^{(0)})}{2\Delta_{\mathrm{d}}c}$, $z^{(0)} = x^{(0)}$

Repeat for $k = 0, 1, 2, \ldots$

$y^{(k)} = (1-\theta_k)x^{(k)} + \theta_k z^{(k)}, \quad \theta_k = \frac{2}{k+2}$

$\mu_k = \alpha\theta_k^2$

$\tilde{u}(y^{(k)}) = \operatorname*{argmax}_{u\in Q_{\mathrm{d}}}\left\{u^T A y^{(k)} - \mu_k d_{\mathrm{d}}(u)\right\}$

$u^{(k)} = (1-\theta_k)u^{(k)} + \theta_k\tilde{u}(y^{(k)})$

if $f(x^{(k)}) - g(u^{(k)}) \leq \epsilon$ **then break**

$\nabla f_{\mu_k}(y^{(k)}) = A^T\tilde{u}(y^{(k)})$

$z^{(k+1)} = P_{Q_{\mathrm{p}}}\left(z_k - \frac{1}{\theta_k L_{\mu_k}}\nabla f_{\mu_k}(y^{(k)})\right)$

$x^{(k+1)} = (1-\theta_k)x^{(k)} + \theta_k z^{(k+1)}$

where $x^{(k)} \in \mathbb{R}^{N\times 1}$. For the algorithm **Restart**, we record the total number of inner iterations k. As primal and dual prox-function we use $d_{\mathrm{p}}(x) = \frac{1}{2}\|x - x^{(0)}\|_2^2$ and $d_{\mathrm{d}}(u) = \frac{1}{2}\|u\|_2^2$ as in [9] ($\hat{u} = 0$). By choosing the center of the primal prox-function as the starting point for the new iterations, we obtain a good initial/warm starting point in each restart [9]. For a fixed accuracy, it was suggested in [9] to use a fixed number of restarts. However, since we sweep over a large range of accuracies it is more appropriate to allow a variable number of restarts. We therefore select γ as suggested in [11].

5.1. Total Variation Denoising

Our first example is the total variation denoising problem [10, 16],

$$\begin{array}{ll}\text{minimize} & \displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n}\|D_{ij}x\|_2 \\ \text{subject to} & \|x - b\|_2 \leq \delta\end{array}$$

where $D_{ij}x$ is an approximation of the gradient at pixel i, j, and m, n is the image dimensions with the number of variables $N = mn$. We observe $b = x_0 + e$ with x_0 the original image and e being i.i.d. Gaussian noise. As initialization we use $x^{(0)} = \bar{x}^{(0)} = b$ and

$$u^{(0)} = \bar{u}^{(0)} = \operatorname*{argmax}_{u\in Q_{\mathrm{d}}} u^T A x^{(0)} - \frac{\epsilon}{2\Delta_{\mathrm{d}}}d_{\mathrm{d}}(u) \tag{8}$$

with $A = [D_{11}, D_{12}, \cdots, D_{mn}]$. For the total variation denoising problem we obtain the simulation results shown in Fig. 1. We observe that the algorithm **Fixed** with fixed μ converges approximately linear $\mathcal{O}(1/\epsilon)$. If we, however, apply **Restart** then the algorithm is faster and the complexity is lower (the slope is closer to that of $\mathcal{O}(1/\sqrt{\epsilon})$ compared to $\mathcal{O}(1/\epsilon)$). The proposed approach **Smooth** with decreasing μ for each iteration converges faster and shows slightly better complexity, approximately $\mathcal{O}(1/\sqrt{\epsilon})$.

5.2. ℓ_1-norm Minimization

For the second example, we will consider the problem of finding a sparse representation of an image b in an overcomplete dictionary B:

$$\begin{array}{ll}\text{minimize} & \|z\|_1 \\ \text{subject to} & \|Bz - b\|_2 \leq \delta\end{array} \tag{9}$$

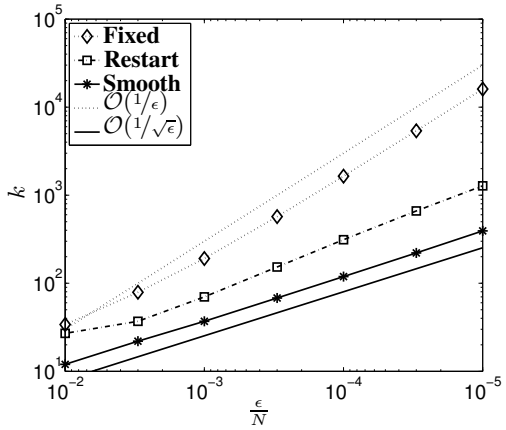

Fig. 1. Simulation results for a total variation denoising example of a noisy image of Lenna (512×512). We report the number of iterations k required to reach the relative accuracy $\frac{\epsilon}{N}$. As a reference, we also show the complexity functions $\mathcal{O}(\frac{1}{\epsilon})$ and $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$.

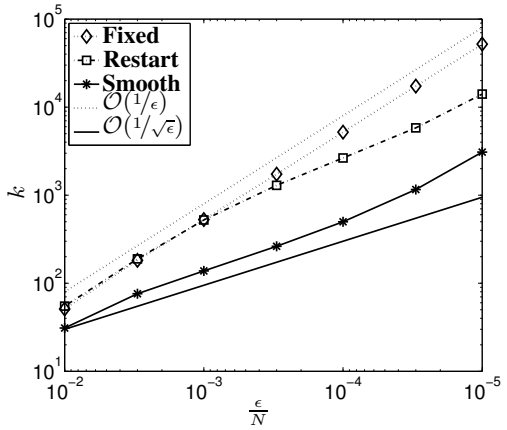

Fig. 2. Simulation results for an ℓ_1-norm minimization example using the image of Lenna (512×512). We report the number of iterations k required to reach the relative accuracy $\frac{\epsilon}{N}$. As a reference, we also show the complexity functions $\mathcal{O}(\frac{1}{\epsilon})$ and $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$.

where $B = [B_1; B_2]$ and B_1 is the 2-dimensional discrete cosine transform and B_2 is a Symlet16 wavelet transform with 3 levels. As shown in [17], the problem (9) can be posed as an equivalent problem with simpler projection constraints

$$\begin{array}{ll} \text{minimize} & \|Wx\|_1 \\ \text{subject to} & \|x_1 - b\|_2 \leq \delta \end{array}$$

$$\text{where} \qquad W = \begin{bmatrix} B_1 & -B_1 B_2^{-1} \\ 0 & I \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

We initialize the algorithms with $x^{(0)} = \bar{x}^{(0)} = [b; 0]$ and $u^{(0)} = \bar{u}^{(0)}$ as in (8) with $A = W$. For this problem we obtain the simulation results shown in Fig. 2, where we again observe that the approach **Fixed** with fixed μ converges approximately linear $\mathcal{O}(1/\epsilon)$. For the **Restart** approach, the convergence rate is closer to $\mathcal{O}(1/\sqrt{\epsilon})$ for high accuracy (small ϵ) but approximately $\mathcal{O}(1/\epsilon)$ for low accuracy. The proposed algorithm **Smooth** with decreasing μ for each iteration converges faster and shows lower complexity than the other two methods. We also generated 100 problems with the vector $b \in \mathbb{R}^{128^2 \times 1}$ being i.i.d. Gaussian. For these simulations we observe similar results as reported in Fig. 2 for the relative convergence speed and complexity.

6. CONCLUSIONS

We presented a new method to speed up the convergence for non-smooth problems using accelerated gradient methods for convex minimization. We provided a proof of convergence, which resulted in a loose bound on the complexity. In fact, practical simulations revealed that the complexity is lower than what the bound suggests. For comparison, we studied and simulated existing methods. The simulations showed that the proposed method has both faster convergence and lower complexity compared to the existing methods.

7. REFERENCES

[1] Yu. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$," *Doklady AN SSSR (translated as Soviet Math. Docl.)*, vol. 269, pp. 543–547, 1983.

[2] Yu. Nesterov, "On an approach to the construction of optimal methods of minimization of smooth convex functions," *Ékonom. i. Mat Metody.*, vol. 24, pp. 509–517, 1988.

[3] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," submitted to SIAM J. Optim., 2008.

[4] Yu. Nesterov, "Smooth minimization of nonsmooth functions," *Math. Prog. Series A*, vol. 103, no. 127-152, 2005.

[5] P. Weiss, G. Aubert, and L. Blanc-Féraud, "Efficient schemes for total variation minimization under constraints in image processing," Tech. Rep., INRIA Sophia Antipolis, France, 2008.

[6] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. on Imaging Sciences*, vol. 2, pp. 183–202, 2009.

[7] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, Dec. 2007.

[8] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for ℓ_1-minimization with application to compressed sensing," *SIAM J. Imaging Sciences*, vol. 1, no. 1, pp. 143–168, 2008.

[9] S. Becker, J. Robin, and E. J. Candés, "NESTA: A fast and accurate first-order method for sparse recovery," 2009, submitted.

[10] J. Dahl, P. C. Hansen, S. H. Jensen, and T. L. Jensen, "Algorithms and software for total variation image reconstruction via first-order methods," *Numer. Algo.*, vol. 53, pp. 67–92, 2010.

[11] A. Gilpin, J. Peña, and T. Sandholm, "First-order algorithm with $O(\ln(1/\epsilon))$ convergence for ϵ-equilibrium in two-person zero-sum games," 2008, 23rd National Conference on Artificial Intelligence (AAAI'08), Chicago, IL.

[12] Yu. Nesterov, "Gradient methods for minimizing composite objective functions," Tech. Rep., CORE Discussion Papers series, Université Catholique de Louvain, Center for Operations Research and Econometrics, 2007, Available http://www.uclouvain.be/en-44660.html.

[13] W. Yin E. Hale and Y. Zhang, "A fixed-point continuation method for ℓ_1-regularization with application to compressed sensing," 2007, CAAM Technical Report TR07-07, Rice University, Houston, Texas.

[14] Yu. Nesterov, *Introductory Lectures on Convex Optimization*, Kluwer Academic Publishers, 2004.

[15] G. Lan and R.D.C. Monteiro, "Iteration-complexity of first-order augmented lagrangian methods for convex programming," submitted.

[16] T. F. Chan and J. Shen, *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*, SIAM, Philadelphia, 2005.

[17] J. Dahl, J. Østergaard, T. L. Jensen, and S. H. Jensen, "An efficient first-order method for ℓ_1 compression of images," in *Proc. of IEEE Inter. Conf. on Acoustics, Speech, and Signal Process.*, 2009, pp. 1009–1012.